Summer 2014

Big data analytics with NYC taxicab data

Xinwu Qian
Purdue University

Graduate School ETD Form 9
(Revised 01/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Xinwu Qian

Entitled
BIG DATA ANALYTICS WITH NYC TAXICAB DATA

For the degree of Master of Science in Engineering

Is approved by the final examining committee:

Dr. Satish V. Ukkusuri

Dr. Fred L. Mannering

Dr. Hubo Cai

To the best of my knowledge and as understood by the student in the *Thesis/Dissertation Agreement. Publication Delay, and Certification/Disclaimer (Graduate School Form 32)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Dr. Satish V. Ukkusuri

Approved by Major Professor(s): _____

_____

Approved by: Dr. Rao Govindaraju 07/29/2014

Head of the Department Graduate Program Date

BIG DATA ANALYTICS WITH NYC TAXICAB DATA

A Thesis

Submitted to the Faculty

of

Purdue University

by

Xinwu Qian

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Engineering

August 2014

Purdue University

West Lafayette, Indiana

For my parents and my beloved.

ACKNOWLEDGEMENTS

I give my sincere gratitude to Prof. Satish Ukkusuri for providing me the precious opportunity to join his research group in 2012. During the past two years, I am greatly influenced by his enthusiasm and dedication in research works and I am highly motivated to conduct excellent research works. He is such a great mentor and also a friend. Also, I would like to thank Xianyuan Zhan and Wenbo Zhang for their direct contributions to part of the thesis. I also need to thank Prof. Fred Mannering and Prof. Hubo Cai for agreeing to serve on my master's committee. Professor Mannering provides valuable guidance on the selection of the proper econometric models and the way to construct the model. Thanks to all of my lab mates - Rodrigo Mesa, Abdul Husain, Binh Luong, Feng Zhu, Pulkit Parikh, Arif Mohaimin Sadri, and Kien Trung Doan for their comments on improving my presentations. Lastly, I must thank my parents, who always stand by my side without any reservation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

o	Trip pick-up location
d	Trip drop-off location
p_i^j	Land use pattern for trip i at location j
d_i	Trip distance for trip i
t_i	Trip travel time for trip i
β	Exponent of power-law distribution
α	Exponent of truncated power-law distribution
λ	Exponent of truncated power-law distribution
ε_{ij}	Error term for the count data model
y_{ij}	Trip amount from ZCTA i to j
p_{ij}	Probability of being in the zero state
$\Gamma(\,)$	Gamma distribution
L_{ij}	Length of link j in the i_{th} ZCTA
n_{ij}	Number of lanes of link j in the i_{th} ZCTA
A_i	Area of ZCTA i
y_i	Total number of taxi pick-ups at ZCTA i
x_{ik}	Independent variables k at ZCTA i
a_{ik}	Coefficient factor for variable x_{ik}
w_{ij}	Weighting factors of ZCTA i on observation j
b	Bandwidth

ABSTRACT

Qian, Xinwu. M.S.E., Purdue University, August 2014. Big Data Analytics with NYC Taxicab Data. Major Professor: Dr. Satish V. Ukkusuri.

In recent years, the fast development of pervasive computing technologies gives us access to a variety of big data sources, such as mobile phone users, bank note movements and twitter posts etc. It provides unprecedented opportunities to conduct urban studies and contributes to the sustainable and efficient development of large cities. While big data is an invaluable treasure, our knowledge is limited due to lack of appropriate tools to analyze the data. Hence, this thesis is mainly focused on demonstrating how taxicab data can be used to study urban travel patterns and explore useful insights.

The first objective of the thesis is to capture the urban dynamics of NYC at the aggregate level. We start by analyzing pick-up and drop-off locations separately and reveal the general trip pattern across the city and the existence of unbalanced trips. The inherent similarities among taxi trips are further investigated using the two-step clustering algorithm. Moreover, human mobility pattern are inferred from the taxi trip displacements and is found to follow two states: an exponential distribution with short trips and a truncated power law distribution for longer trips. The findings indicate that taxi trips can be viewed as reasonable proxy for human mobility, but the result is subject

to the intervention of internal and external factors such as trip fare, urban form and land use functionality.

The second objective of the thesis is to investigate the factors that influence the urban taxi demand from two aspects. The first subtask is to understand the intra-city movement by taxicabs. The Zero-Inflated Negative Binomial model (ZINB) is implemented considering the nature of count data and the excessive zeroes. The results indicate that distance, land use, demographic and socioeconomic variables all have significant impact on the intracity taxi trips. The ZINB model illustrates why there is an absence of taxi trips between a particular OD pair. The other subtask is to analyze the spatial variation of taxi ridership. Except for variables in the first subtask, access to other transit is also introduced. The geographically weighted regression (GWR) is used to model the urban taxi ridership and visualize the spatial variation of independent variables. The results suggest that GWR model outperforms the ordinary least square (OLS) method in both goodness of fit and explanatory power. Additionally, the urban form is found to greatly influence the calibration of independent variables and failing to account for spatial non-stationary effects may lead to biased estimations. The results of both subtasks provide an in-depth understanding on the spatiotemporal variation of urban taxi demand and may serve as guidance for predicting taxi demand, developing proper regulations for the taxi industry and drawing up urban plans.

CHAPTER 1. INTRODUCTION

1.1 Urban Dynamics

The rapid urbanization process gives birth to megacities such as Tokyo, Shanghai and New York City (NYC). Megacity creates tremendous job opportunities and boosts the economy growth. However, due to the influx of people and intensive activities, megacity also incurs great challenges such as traffic congestions, environmental concerns and crimes.

Urban dynamics represents the spatiotemporal principles followed by urban functioning evolvements (Sun et al., 2013). Understanding urban dynamics helps to capture the pulse of urban activities and principles of human movement, which in return provides potential solutions to address the management and operation issues of large cities. First, the temporal and spatial distribution of people in urban area is essential to urban planning. For example, the congestion level of the traffic network can be obtained from the urban dynamics and proper strategies such as building new roads or proposing new subway line can be made. Second, urban dynamics may contribute to policy making. Given the same example, policies can be assigned to ease the congestion instead of constructing new facilities. Furthermore, urban dynamics is beneficial towards system management. The movement pattern of people during large social events such as Super

Bowl game night can be obtained from urban dynamics. Such pattern should be considered as valuable experience for the management of other large activities.

In the past few decades, efforts have been made in modeling and simulating urban dynamics using data from transportation systems (Batty and Xie, 1994; Batty et al., 1999; Harris, 1985). However, as a complex system, it is always hard to describe a city accurately from mathematical models. The difficulties are mainly resulted from the heterogeneity of human behavior, the urban forms and the effect of geographical boundary. Therefore, there is a need to develop more efficient tool that allow us to understand the urban dynamics from transportation systems.

1.2 NYC Taxicab

Taxicabs, especially in large cities, assume indispensable functionality in urban transport system. It complements other public transport modes in terms of the flexible door-to-door service and 24-7 operation hours. It also carries a large amount of passengers daily.

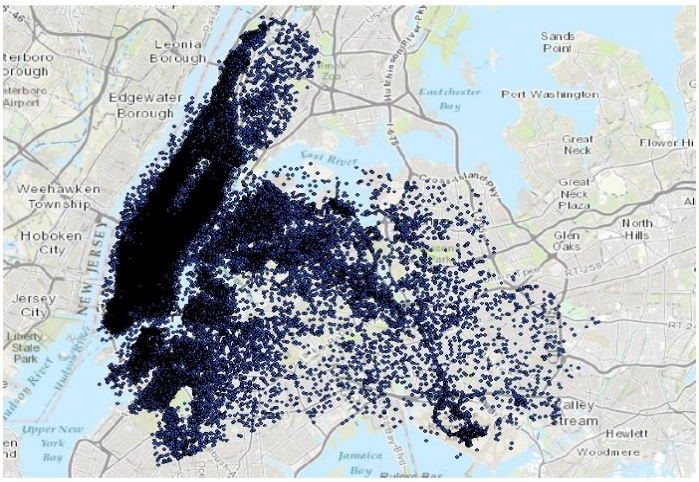


Figure 1.1 The distribution of taxi OD across NYC

The yellow cab are often regarded as the symbol of NYC. Figure 1.1 presents the spatial distribution of taxi pick-ups and drop-offs on a particular day. The service area covers five boroughs in NYC, including Manhattan, Brooklyn, Bronx, Queens and State Island. According to the New York City Taxi &Limousine Commission (NYCTLC, 2012), by the end of 2012, there are more than 13,000 medallion taxicabs which makes NYC the largest taxi market in North American. Moreover, around 450,000 taxi trips are made every day, transporting over 800,000 passengers around the city. This contributes to ten percent of the total passengers served by public transportation. Since most taxis in NYC are equipped with GPS devices, they work as "floating" vehicles around the city providing plentiful amount of useful trip information. The data collected from the NYC taxicab is therefore utilized in the research for further urban studies.

1.3 Motivations

In the era of big data, we have access to various data sources. In urban areas, the abundant information from different data sources contributes to a comprehensive understanding of activity patterns. Several pioneering studies have implemented various data sources to reveal urban activity participation and individual mobility patterns. For example, the traces data is obtained from the mobile phone users to reveal the activity dynamics in urban area (Ratti et al., 2006; Reades and Calabrese, 2007; Calabrese et al., 2013). The distribution of bank notes is used to study human mobility (Brockmann et al., 2006). Hasan et.al (2013) examined both aggregate and individual activity patterns from social media check-in data. The data driven methods are gaining popularity, however, a critical question to answer is how much can we learn from the data and to what extent we

are able to interpret it. In fact, these data sources are suffered from several drawbacks. One drawback in common is that the activity pattern are inferred rather than directly carried in the data.

In contrast with other data sources, the taxi trip data enjoys the merit of sufficient temporal and spatial coverage due to the large passenger volume and 24-7 operation hours. Moreover, it is more interpretable since the geographical locations obtained are straightforward representation of human activities. Therefore, it is an advantageous data source for urban studies and has already received great attentions from researches. The pick-up and drop-off locations are processed with data mining and clustering algorithms to reveal urban activity patterns such as hotspots information for taxi drivers (Chang et al., 2008; Yuan et al., 2011; Li et al., 2011) and for land use inference (Pan et al., 2013; Liu et al., 2012). However, urban dynamics are understood superficially if pick-up/drop-off locations are only analyzed separately. Taxicab provides door-to-door service and is often used as a direct commuting tool. Therefore, the joint of pick-up and drop-off location builds up a direct bridge between origin and destination and can aggregately reveal the underlying connections among separated urban places. Moreover, since taxi trip is a specific form of human movement, the taxi trip data can also be used to disclose the uniformity of human mobility in large cities (Peng et al., 2012; Liang et al., 2012; Jiang et al., 2009). However, discrepancies of the result are observed among limited works. As a special case of human mobility, the displacement of taxi trips is restrained by the trip expenses and more importantly, the structure of a city. Therefore the relationship between taxi trips and human mobility requires further investigation.

The second motivation arises from the availability of geographical databases. The emphasis of most big data works is associated with pattern recognition. Nevertheless, simply retrieving information from the big data alone is far from enough to tackle big challenges in large cities. Compared with traditional data collected from surveys and questionnaires, the pervasive computing devices are able to collect abundant data in an efficient and accurate manner. The combination of big data with statistical and econometric analysis therefore helps to address the issues of clarifying the determinant factors for the urban dynamics. The geographical database enables a geographical representation of demographics, socioeconomic variables and facility locations etc. As a result, it is worthwhile to investigate the combination use of geographical database and the taxi trip coordinates to explore in-depth reasons behind urban issues.

1.4 Research Objectives

There are three main objectives in this research. The first objective is to reveal urban activity patterns and explore insights from the taxicab data. Since taxicab data contains the temporal and spatial attributes of taxi origins and destinations, it can be utilized to analyze and visualize the activity dynamics across the city. Besides, the connection of urban areas can be constructed by understanding the inherent similarities of taxi trips. Moreover, as a special form of urban movement, it is meaningful to retrieve the human mobility from taxi trips.

The second objective is to characterize the factors that are related to intra-city movement. The intra-city movement is of short travel distance, high trip frequency and various trip purposes. More sustainable transportation control and planning strategies can

be developed by understanding the driving forces that migrates people around the city. While inter-city and inter-regional human movements are widely examined, it is difficult to study the intra-city movement due to the intricate flow variation and lack of data. Therefore, we tend to investigate the determinant factors for intra-city human migration from the NYC geographical database and taxi trip data.

The last objective is to explore the spatial variation of taxi ridership across the city. Understanding taxi ridership is essential to predict taxi demand, design proper regulation for taxi industry and plan for new constructions in urban area. Demographics, socioeconomics and the existence of other transportation modes are all associated with the taxi ridership. However, due to the impact of urban form and functionality, it is still unclear how effects of these variables varying over space. Hence, the spatial analysis will extend our knowledge on the urban taxi ridership.

1.5 Literature Review

1.5.1 Research on Big Data in Urban Analysis

Several pioneering studies mainly focused on mobile phone data to reveal basic urban activity and individual mobility patterns. A case study in Milan successfully discovered the urban spatial and temporal variations of activity intensity (Ratti et al., 2006) by using location-based services (LBS). The study area is an approximately square in the size of 400 square kilometers around the center of Milan. Individual locations are collected and mapped onto actual maps and urban dynamics is presented from aggregated level by analyzing the temporal and geographical feature of phone call density. The intensity of activity locations is further used to locate hot spots and identify city structure

by analyzing spatiotemporal signatures of Erlang data, which is a measure of network bandwidth usage and an Erlang is one person-hour usage of mobile phone (Reades and Calabrese, 2007). The research suggests the feasibility to build a real-time representation of dynamics at city-region scale from mobile phone network. Furthermore, the results are found to be consistent with theoretical researches on impacts of telecommunication on urban behaviors. Individual mobility is also studied by using mobile phone traces (Calabrese et al., 2013). A multivariate regression model is proposed to study the intra-urban spatial pattern of individual and vehicular mobility with respect to population density, land use mix, street network layout, and accessibility. The result indicates that mobile phone traces are reasonable proxy for individual mobility. From individual perspective, a highly regulated human mobility pattern is revealed from 100,000 mobile phone users' trajectories (Gonzalez et al., 2008). Contrary to the conventional idea that human trajectories are random, high regularities both temporally and spatially are observed for human movements. It is found that people are very likely to return to frequently visited places and human behaviors follow reproducible patterns.

In recent years, taxi data has also attracted broad attention in urban studies. Data mining approach is implemented to predict hotspots from taxi data (Chang et al., 2008). Trip request records are filtered based on predefined context. K-means, DBSCAN, and agglomerative hierarchical clustering algorithms are used to classify the retrieved data and thus hotspots are specified. While it is hard to get the detail use of land use mix, a taxi traces based methodology is proposed to recognize the social functionality of a particular region (Pan et al., 2013). A predictive model is built to estimate the vacant trips given proper inputs such as time features and weather conditions (Phithakkitnukoon,

2010). The model mainly consists of inference engine and error-based learning, where the inference engine is based on maximizing a posterior (MAP) method and a uniform weighted function is used for recently errors in case of possible changes in the pattern.

1.5.2 Research on modeling human migration

Human mobility is an important indicator of the functional relationship among places (Boyle et al., 1998). Considering the scheme of travelling, it includes any movement over the space such as journey-to-work trip, migration, and the movement of any commodities. It is noted that all kind of trips will give rise to spatial interactions among places and gravity model and its variations are widely applied to model such interactions (Haynes and Fotheringham, 1984). Most researchers studying human movement focus on intraregional migration and interregional migration. The inter-municipal study in Japan reveals two major migration patterns in late 1980s and agglomeration effect and competing effect are observed by using competing destination model (Yano et al., 2000). The population migration within United Kingdom is mapped by using 2001 UK census data (Rae, 2009). The nationwide migration pattern and the in and out flow pattern at particular cities are further discussed. A modified gravity model is constructed to analyze the internal migration in Russia considering unemployment rate, poverty, natural resources, and socio-political conflict etc. (Andrienko and Guriev, 2004). The result suggests that migrants will be attracted by improving living standards, creating jobs and improving public goods provision. Also, the amount of migration flow is very sensitive to the distance.

For interregional movement, the influence of political, economic and demographic factors are examined on the size of migration flows to North America (Karemera et al., 2000). An extension of canonical gravity model is implemented and the result indicates that the increasing of population associated with origin and destination as well as the economic development level have significant impacts. The return of migrants in Eastern Europe is shown to have positive impact on the productivity level of the source country (Leon-Ledesma and Piracha, 2004).

However, a potential problem that may arise from both intraregional and interregional migration is the significant amount of zero flows between origins and destinations (Nakaya, 2001; Linders and Groot, 2006; Lesage et al., 2007; Deng and Athanasopulos, 2011; Mata and Llano-Verduras, 2012). The general approach using gravity model is to take the transformation into log-scale and use Ordinary Least Square (OLS) estimation. The issue associated with the approach is the infeasible logarithm value of zero (Anderson and Wincoop, 2004). Methods are proposed to deal with the problem such as deleting zero values or add a small amount of flow to avoid zero flows (Linders and Groot, 2006). An extension of Tobit model (Tobin, 1969) called Bivariate normally distributed Probit Regression (BPR) model is constructed (Bikker, 1992) which can be estimated easily and account for zero flows. Instead of log-normality assumption of gravity model, a Poisson specification of the gravity model is proposed which is able to estimate zero flows (Silva and Tenreyro, 2006). Nevertheless, the mean and the variance of OD flows are usually not equal thus violating the underlying assumption of Poisson distribution. To overcome the drawback, negative binomial and zero-inflated

negative binomial specifications of gravity model are built (Burger et al., 2009) which are able to deal with the overdispersion pattern and the existence of excessive zeroes.

1.5.3 Research on Ridership Analysis

The importance of ridership forecasting is self-evident as it provides the future demand for stakeholder to make suitable policies on planning. In a sense, the demographics such as population and employment are viewed as proxy for transportation demand (Mcnally, 2008). On the other hand, the demand can also be examined from economics theory as a utility function with respect to travel time, costs, and level of service etc. (Ben-Akiva and Lerman, 1985; Kanafani, 1983). In order to forecast the demand of public transportation, extensive studies are conducted to understand the factors associated with transit ridership. From the geographical point of view, density, diversity and design of built environment is thought to have a close relationship with travel demand (Cervero and Kockelman, 1997). Transit service has a higher chance to be used if more people live or work nearby (Murray et al., 1998). The impact of demographics on ridership is known to be significant. In general, the poor, the elderly, the minorities and women have a high dependency on mass transit (Pucher et al., 1981). However, the dependency may vary as substantially differences are observed between women and men (Cristaldi, 2005). From economic perspective, the gasoline price is found to have a substantial impact on bus ridership with elasticity equals to 0.42 (Agthe and Billings, 1978). Though bus ridership is expected to increase with higher gasoline price, it is suggested that the additional fare revenues may not be able to cover the extra fuel expenses of transit systems (Mattson, 2008).

In order to obtain an accurate estimation, intensive approaches are made to forecast transit ridership. Traditional approaches usually use multiple regression model to construct the relationship between ridership and explanatory variables. However, one drawback of such model is the ignorance of spatial interactions which are believed to have significant impact on transportation demand. For example, people living in a buffer 500 to 1000 meters away from a railway station tend to use rail services are 20% less likely to use railway service compared with people living less than 500 meters from stations (Keijer and Rietveld, 2000). Moreover, the patronage of bus is observed to decline with increasing walking distance and a distance-decay weighted regression is built to estimate the bus ridership (Gutierrez et al., 2011). Apart from the distance, which represents the geographical accessibility, the explanatory variables may have spatial variations as well. One of the preliminary works uses geographically weighted regression (GWR) model to characterize the effects of independent variables varying across space. Though important, there are still very few literatures that captures the spatial interactions when forecasting transit ridership.

1.6 General Framework

The thesis aims at providing a platform on how to analyze large scale taxi data to study urban taxi travel patterns. The taxicab data from NYCTLC is utilized to characterize the spatial and temporal dynamics of human movements in urban areas and model the movement and ridership. The main components of the thesis are presented below:

1. Data processing. The raw taxi data is processed to filter out erroneous information and generate the data set matching up the study boundary. Then data mapping is conducted based on the longitude and latitude for further analysis.
2. Pattern recognition. The overall trip distribution is presented first and five hot spots are selected by trip ranking and land use information. The temporal variation of trip patterns at hot spots and the corresponding implications are discussed in detail. Furthermore, the existence of unbalanced taxi trip pattern is also revealed.
3. Trip classification. A two-step clustering algorithm is implemented to classify trips based on travel distance, trip starting time, geographical location, and land use types at origins and destinations. Seven distinct clusters are specified and their trip starting time and trip distance distributions are discussed.
4. Mobility analysis. Trip distances are partitioned into groups and scatted plot indicates that the mobility pattern of taxi trips follows a heavy-tailed distribution under log scale. Moreover, it is inferred from the plot a two regimes for mobility of taxi trips: decision-making process of making taxi trips and a truncated-power law distribution after a distance threshold.
5. Modeling intra-city movement. The trip frequency is divided into three time periods for each day of the week. Based on the highly skewed distribution of OD frequency and the count data nature of trips, 21 Zero-Inflated Negative Binomial regression models are constructed to explore the influence of demographics and socioeconomic factors on intra-city movement.

6. Taxi ridership analysis. With a concentration in built environment, the ridership of taxicabs is analyzed with a Geographically Weighted Regression (GWR) model. Variables are designed to characterize the accessibility to subway and jobs. The results serve as an ideal reference for taxi market forecast and taxi fleet regulation.

The thesis is organized as follows: Chapter 2 describes the data available and the process of data cleaning and mapping. Chapter 3 presents a comprehensive study on urban dynamics from taxi data, starting from the general trip distribution and delving into details of hot spots dynamics, trip imbalance, classification of taxi trips, and taxi mobility. Chapter 4 analyzes the distribution of OD trip frequency and constructs the econometric models to understand intra-city movements. Chapter 5 discusses the framework and advantage of GWR and presents the estimation results for taxi ridership. Conclusion and remarks are given in Chapter 6.

CHAPTER 2. DATA

2.1 Taxi Data

The taxi trip data used in this research is collected by New York City Taxi &Limousine Commission (NYCTLC) from December, 2008 to January, 2010. About 350,000 to 550,000 daily trips are recorded and the annual trip distribution is plotted in Figure 2.1. A reproducible and stable pattern is observed for weekly trips over the year and inflections are detected during holidays such as Thanks Giving and Christmas. The dataset contains detailed trip information, including the pick-up and drop-off timestamps and locations, the number of passengers onboard, the travel distance and the trip expense. Complete trip trajectories are not available due to privacy concerns.

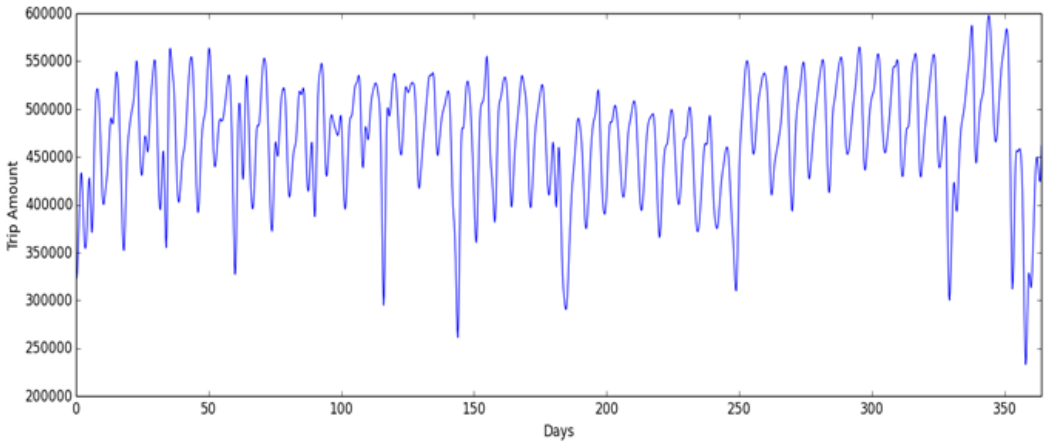


Figure 2.1 2009 Taxi Daily Trip Distribution

Due to the overall repeatable and stable pattern over the week, one week trip data from October 5th to October 11th are extracted for further analysis. The time period contains no major social events and has the temperature at the average level. The statistics of the one week data is presented in Table 2.1. The data is first processed before further analysis. Erroneous trip records are firstly eliminated, such as trips with zero travel distance or fare less than the initial price. The Staten Island is removed from the study area due to very few trip observed. Then all pick-up and drop-off locations are coupled with geography map to clean out trips outside the study area. There are over 3 million trips after processing and all trips are tagged with the overlaid census tract ID, zip code and land use type.

Table 2.1 Taxi Data Statistics

Date	Number of Trips Recorded	Number of Trips after Cleaning
10.5.2009	431,828	428,553
10.6.2009	467,649	464,273
10.7.2009	492,914	488,895
10.8.2009	517,079	512,781
10.9.2009	536,039	531,965
10.10.2009	532,179	528,032
10.11.2009	454,573	451,059

2.2 Geographical File

In addition to taxi data, census tract and ZIP Code Tabulation Areas (ZCTAs) are also implemented as the geography representation of the study area. The census tracts are

extracted from the census tract area file provide in TransCAD[1]. On the basis of spatial distribution of taxi trips, 2211 census tracts are selected to be the study area, which cover Manhattan, Bronx, Queens, Brooklyn, Long Island, and a small portion of New Jersey. The area covered by census tracts and ZCTAs are presented in Figure 2.2. The ZCTAs shape file is extracted from NYC Geodatabase[2]. There are 210 ZCTAs under the same geographical boundary as the census tracts. The ZCTAs is processed following two steps: (1) the sub-ZCTAs are merged with the parent ZCTA as presented in Figure 2.3 and (2) the ZCTAs with zero population are removed. The amount of ZCTAs reduces to 168 after processing. The census tracts is used as the basis for pattern recognition in order to get finer estimation while ZCTA is used for modeling taxicab trips considering the number of OD pairs is of quadratic size to the number of areas.

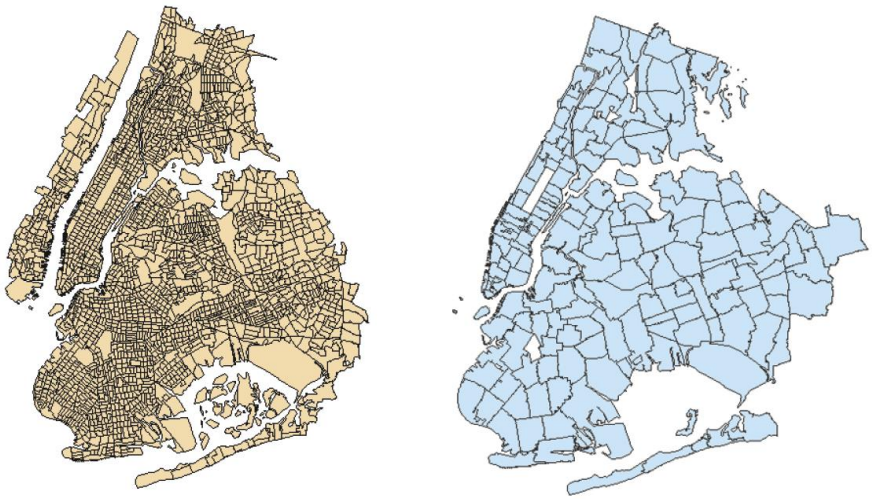


Figure 2.2 Census Tract Map (Left) and ZCTA Map (Right) for the Study Area

[1] TransCAD, a Transportation Planning GIS software by Caliper Corporation.

[2] NYC Geodatabase: http://www.baruch.cuny.edu/geoportal/nyc_gdb/

Moreover, land use information, 2010 census survey, 2009 American Community Survey (ACS) and several geographies of facility locations are also utilized. The land use map implemented in the study is obtained from New York City Department of City Planning (NYCDCP), which divides the city into three fundamental zoning districts: commercial (C), residential (R) and manufacturing (M). The last three types are further categorized from low density to high density. The census survey, ACS and facility locations are part of the NYC Geodatabase. It provides the flexibility to couple the important social statistics with the geographical file so that we can explore the important factors related to the travel pattern of taxicabs in NYC.

Figure 2.3 Example of sub-ZCTAs

CHAPTER 3. PATTERNS OF URBAN DYNAMICS

3.1 Overall Pattern

In this section, patterns of urban activity participation are examined from the arrival and departure dynamics of taxi trips. NYC is one of the busiest cities in the world. Around 5.7 million passengers move around the city during the study period, generating more than 3.4 million taxi trips. The pick-up and drop-off location of all trips are aggregated at the census tract level based on the geographical coordinates and the overall geographical distributions of taxi trips are visualized in Figure 3.1.

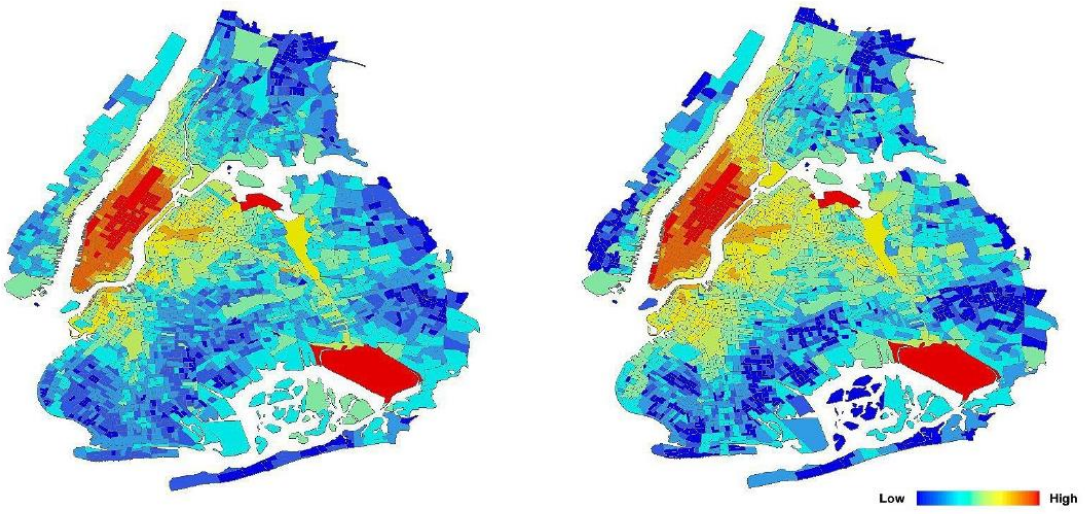


Figure 3.1 Aggregated Weekly Trip Distribution Plot

The most appealing observation is that both trip origins and trip destinations exhibit highly centralized distribution towards Manhattan area. The result is not

surprising since Manhattan serves as the business center of NYC. The number of trips drops immensely with the increasing distance to the city center, which reflects the typical sprawl of urban forms. While most places far from Manhattan have very low amount of trips, patterns at LaGuardia airport (LGA) and John F. Kennedy international airport (JFK) are entirely different. Figure 3.2 shows statistics of the regions where most trips took place. Approximately 90% of total trips are associated with Manhattan area. While a majority of the trips are congregated at midtown Manhattan and lower Manhattan, the upper Manhattan area is apparently less preferred by both passengers and drivers.

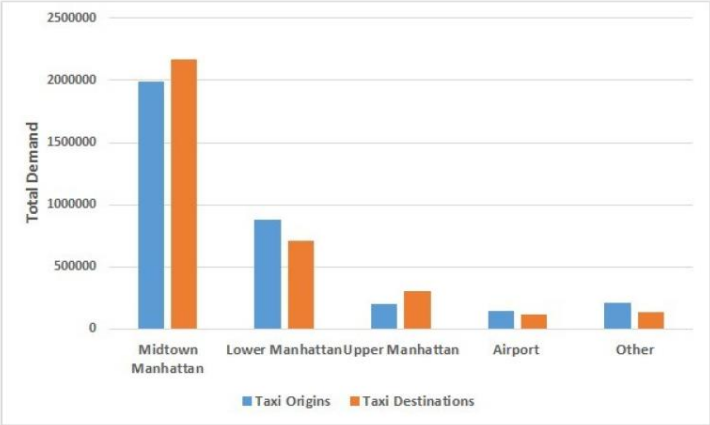


Figure 3.2 Weekly Trip Amount by Areas

3.2 Hot Spots Analysis

The hotspots represent the most frequent visited places in a city and usually come up with great activity intensity. The analysis of hotspots dynamics helps to understand the urban functionality in depth. By ranking total trip frequencies, most popular places are identified and five specific tracts are selected which cover the LGA, JFK, Penn Station, Central Park and the Fifth Avenue (the segment between 49th street and 56th street).

Each individual hotspot has indispensable functionality including transportation terminals (with different purposes), recreational place and commercial area. To analyze the dynamics at hotspots, the temporal patterns across the week are plotted in Figure 3.3.

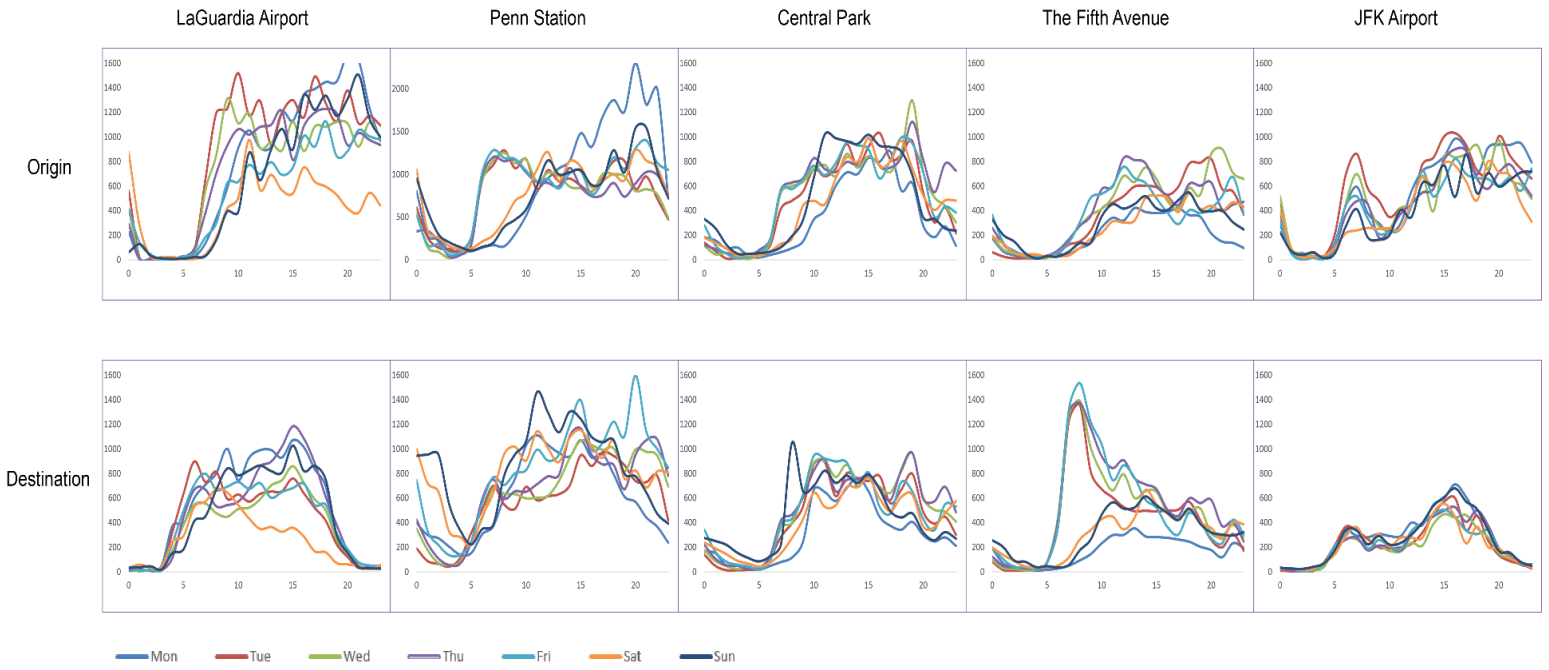


Figure 3.3 Temporal Patterns of Taxi Trips at Five Hot Spots in NYC.
(X-axis: Time of the Day; Y-axis: Trip Frequency)

Penn Station and Madison Square Garden locate in the census tract where the greatest number of taxi trips is generated. Penn Station is not only the terminal for Amtrak trains, but also serves as the connection station for multiple subway lines. According to the morning arrival (trip origin) and evening departure (trip destination) peaks in weekdays, taxicab is very likely to function as the last and first mile transportation. Over the weekend, most arrivals and departures take place within daytime and at night. The pattern coincides with the functionality of Madison Square Garden as an amusement place and a number of hotels nearby.

Trip patterns at airports are distinct from that in central part. For both airports, while arrival curves are comparatively smooth, departure curves are heavily fluctuated due to the periodical entry of flights. Besides, the intrinsic differences between the two airports are also disclosed from trip dynamics. Due to the effect of travel distance, the trip amount at JFK is significant lower than that at LGA. Secondly, since LGA are mainly used for domestic flights, the apparent morning peaks for flight arrivals during weekdays and the drop of trip amount on weekends. Moreover, as an airport mainly for international flights, JFK has more arrivals in the afternoon and the pattern is surprisingly consistent over the week. The result suggests that, during the week, the trip purpose is stable for international flights but changed drastically for domestic flights.

The Central Park is a recreational place. It occupies a larger area compared with other census tracts which contribute to its trip frequency. The comparison of trip dynamics between at the Central Park and at the Fifth Avenue perfectly interprets the functionality of corresponding land use attributes. The Fifth Avenue is a remarkable business street at midtown Manhattan and morning taxi arrival and evening taxi departure

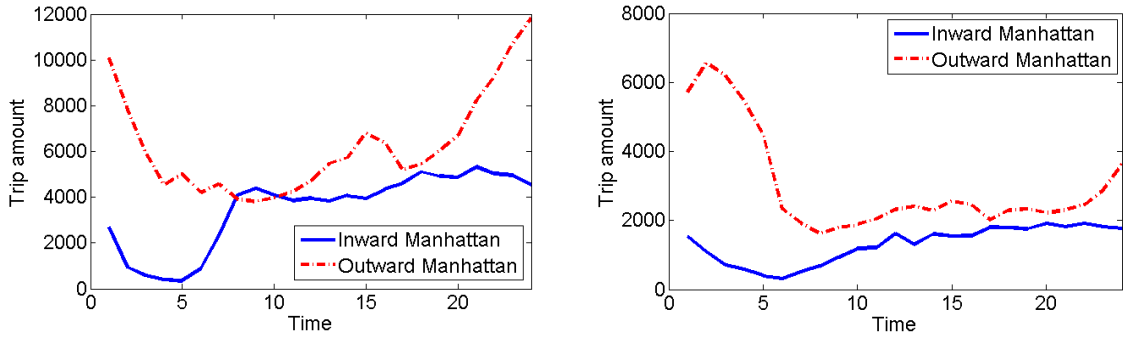
peaks are unsurprisingly retrieved. Reversely, due to the large portion of residential areas around the Central Park, most departures take place in the morning and majority of taxi arrivals are observed during evening rush hours.

3.3 Unbalanced Taxi Trip

Except for being able to capture activity dynamics at hotspots, the data also carries implicit yet significant insights such as the existence of unbalanced taxi trip. The number of taxi trips is closely associated with the level of economic development and the variation of urban functionality. Due to concerns such as trip margins and safety issues, taxi drivers usually have their preferred destinations, which eventually lead to the pattern of geographical discrimination. For example, taxi drivers may be unwilling to make trips to destinations where it is hardly possible to find potential passengers. The second type of unbalanced taxi trips are usually caused by sudden fluctuations in passenger demand. While the supply of taxis is fixed, the influx of commuters during peak hours makes it extremely hard to hail a vacant taxi.

From the overall spatial distribution, we observe a tremendous centrality of taxi trips at the developed Manhattan area. The great trip density suggests the easiness of finding passengers in Manhattan and stickiness of drivers to Manhattan area. As a result, we start looking into phenomenon and plot the temporal distributions for trips inwards and outwards Manhattan in Figure 3.4. Although twisting during daytime, the overall pattern turns out to be stable and balanced. However, when time goes to late night, we surprisingly witness an enormous gap: the highest amount of outward trips and the lowest amount of inward trips take place simultaneously. People may stay at Manhattan very late

for entertainments and relaxations, while buses and metros having a reduced accessibility at the time. As taxi becomes very popular at a late time, drivers may refuse to leave Manhattan as they have to run the risk of returning empty. Hence, the unbalanced trip pattern implies the existence of geographical discrimination and a reduced level of service for taxi industry.



Weekday Weekend

Figure 3.4 Inward/Outward Manhattan Unbalanced Trips

In order to reveal the unbalanced condition inbound Manhattan, we extract only weekday trips and spatial distributions of trip origins and destinations are presented in Figure 3.5. Three typical time intervals are selected which cover off-peaks and morning and evening rush hours. Both morning peak and evening peak display eminent differences between trip origins and destinations and their patterns appear to be symmetric. Moreover, trips are found to be unbalanced with notable geographic characteristics. The northeastern part of midtown Manhattan is a large residential area and the midtown is mainly covered by commercial floors. As a result, most taxi trips inflow into midtown during morning peak and dissipate from the center area in the evening.

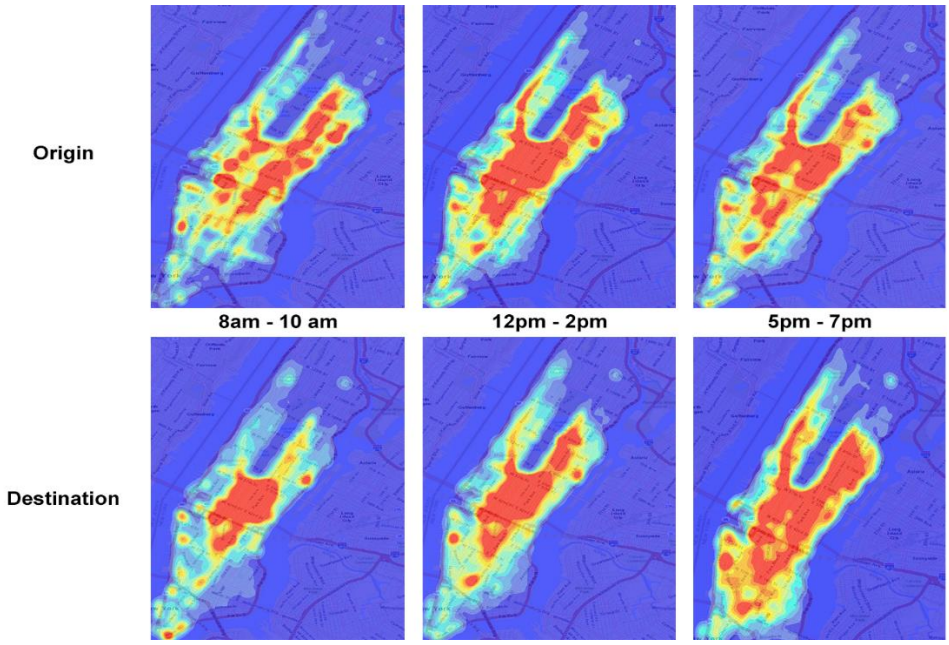


Figure 3.5 Trip Density Plot Inbound Manhattan

The existence of unbalanced taxi trips suggests an imminent need of designing policies to mitigate negative impacts. An additional fee can be charged or a subsidy can be assigned for trips outward Manhattan only after midnight as taxi drives are less likely to leave Manhattan at that time. Moreover, since morning and evening trips have distinct origins and destinations, the shuttle service following the direction of human migration should be to be effective. It can narrow the demand-supply gap of taxi service and reduce congestion at the same time.

3.4 Trip Classification

3.4.1 Two-Step Clustering Algorithm

It is recognized that dynamics of trip origins and destinations are largely influenced by the geographical location, land use pattern and functionality of a particular

place. Moreover, unlike other public transportation modes, the door-to-door service of taxicab builds up the straightforward connection between trip origin and destination. Therefore, how different urban areas are related can be understood by exploring the inherent similarities of taxi trips.

Clustering algorithms are widely used to classify individual cases in large database into homogeneous groups. Considering spatial and temporal characteristics of taxi trips, each piece of taxi trip x_i can be represented as an eight dimensional tuple which takes the form:

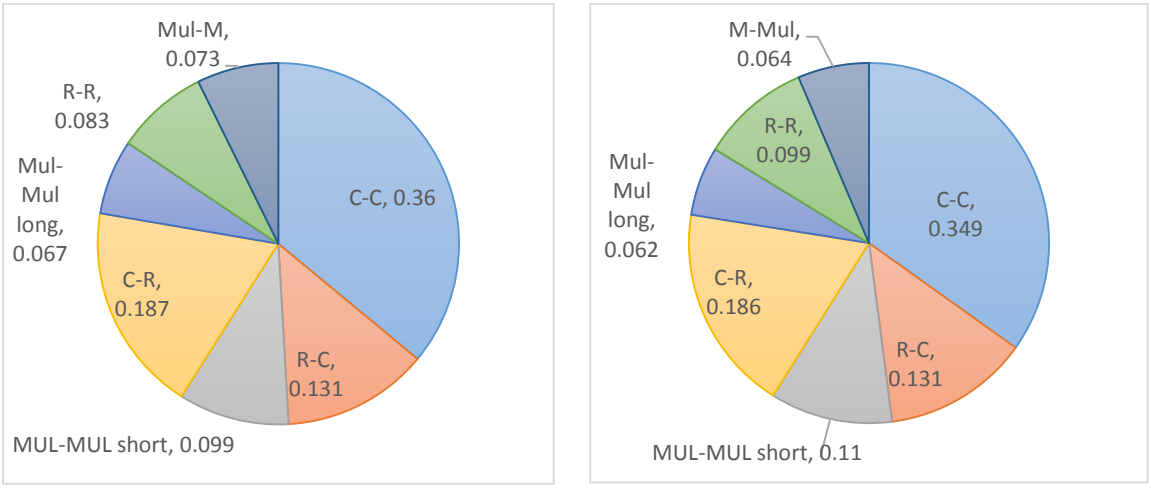
$$x_i = \left(lat_i^o, long_i^o, lat_i^d, long_i^d, p_i^o, p_i^d, d_i, t_i\right) \tag{3.1}$$

Where o, d represent the trip origin and destination respectively, lat and $long$ are the latitude and longitude of trip locations, p refers to the land use attribute, d is the trip distance and t stands for the trip starting time. The clustering problem cannot be tackled by popular approaches such as k-means and DBSACN due to the presence of categorical variables (land use attribute).

Alternatively, the two-step clustering algorithm (Chiu et al., 2001) is implemented to address the mixed variable clustering problem following two stages. The first stage is a pre-clustering approach which uses a sequential clustering method to generate initial sub-clusters. The second stage uses the agglomerative hierarchical approach which processes the sub-clusters from in the first stage recursively. The number of clusters is determined automatically by comparing BIC values. For interested readers, the detailed description for each step of the algorithm can be referred to SPSS manual (SPSS, 2001)

3.4.2 Clustering Results

An overview of the clustering result is presented in Figure 3.6. For both weekday and weekend taxi trips, the exactly same configuration with 7 distinct trip groups is obtained. Moreover, the percentage for the same cluster is pretty close. We name each cluster by its land use feature accordingly, including C-C, R-C, C-R, R-R, Mul (Mixed land use type)-Mul-S (short trip distance), Mul-Mul-L (long trip distance), and Mul-M trips. To better understand the characteristics of each cluster, the spatial distributions of trip origins and destinations in each cluster on weekdays are visualized in Figure 3.7.



Weekday Weekend

Figure 3.6 Clustering Results for Weekday and Weekend

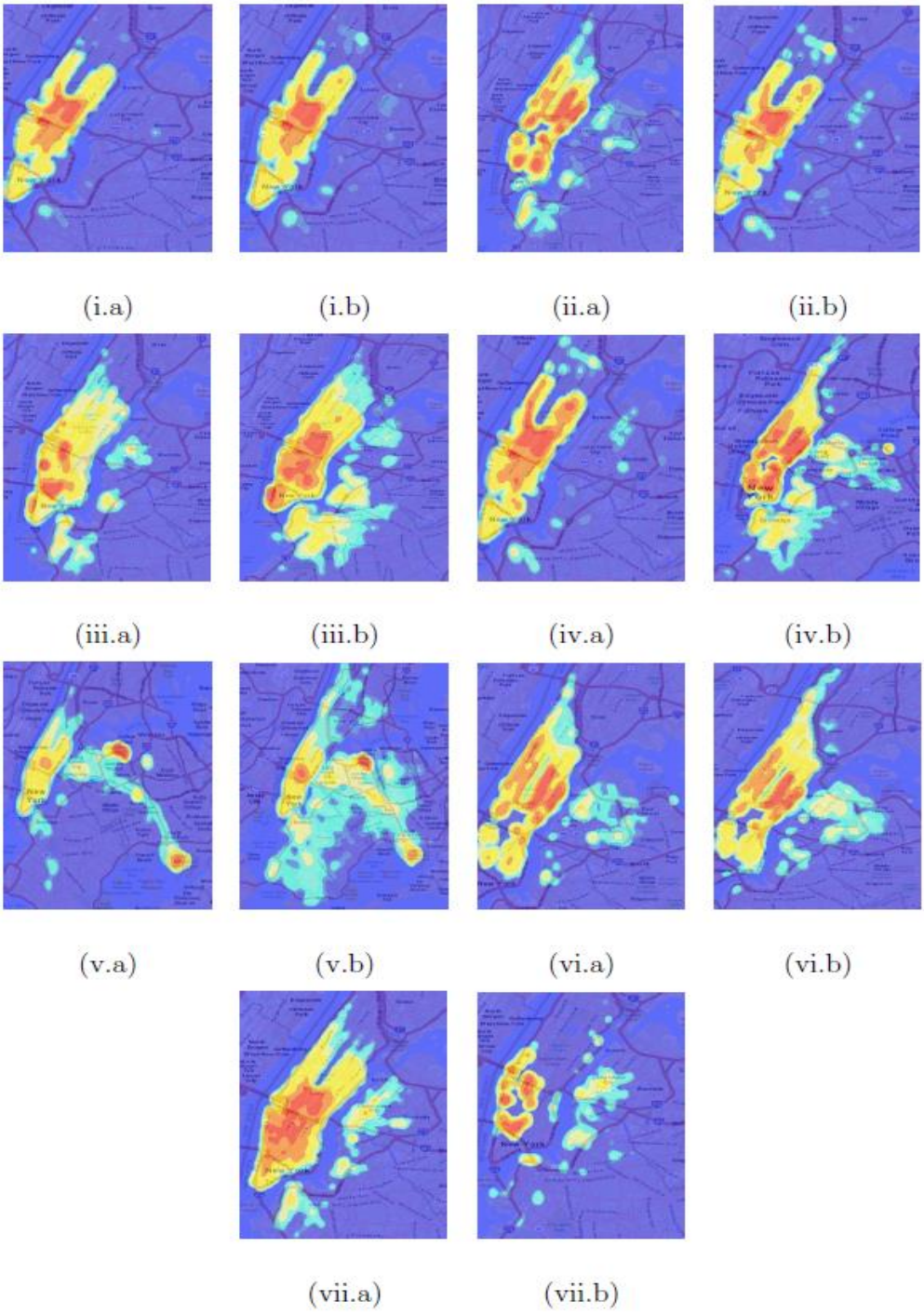


Figure 3.7 Spatial Density Plot of Cluster Origins and Destinations
(i: C-C trip; ii: R-C trip; iii: Mul-Mul short trip; iv: C-R trip; v: Mul-Mul long trip; vi: R-R trip; vii: Mul-M trip; a for origin and b for destination; density increases from blue to red)

In general, C-C trips contribute over one-third (36.0% for weekday and 34.9% for weekend) of the total taxi trips in NYC. Further, there are another 30% of trips that are associated with commercial area (with either origin or destination in commercial area). This suggests the significant impact of land use pattern, especially commercial floors, on the amount of taxi trips. More specifically, commercial areas where trip originated from and arrived at cover the entire midtown and lower Manhattan. As a result, it is believed that most activities and functionalities of the city are concentrated in these places. Viewing the distribution of residential related trips, one can tell that there are considerable amount of people living on the peripheral area and they are connected to the city center by taxicab.

We also plot distributions of travel distance and trip starting time as important attributes for each cluster in Figure 3.8. Apparently, the distance distribution suggests that taxi trips are heavily used for short-range travel, especially for trips less than 5 miles. Such pattern is mainly determined by the urban structure of NYC, as majority activities and functional places are agglomerated in a small area. While the distance distribution is stable over the week, there are prominent discrepancies observed for trip starting time between weekday and weekend. Firstly, all clusters except C-R and Mul-Mul-L trips have morning and evening peaks, reflecting that taxicabs are heavily used for work commuting in urban areas. Secondly, the temporal pattern of most urban activity is shifted from daytime to late night, as the trip intensity remains at a high level until 3 am. However, if we look at the kernel density estimation result for recreation and shopping activity category, we will find a completely different pattern, that the activity seems to be stationary. For the recreation activity, the activity center emerges from the afternoon in

the northwest side of the center Midtown Manhattan, then becomes prominent during nightfall, and finally fades out in the late evening. For the shopping activity, there are several activity centers: two major activity centers located in the central part of Midtown Manhattan, and two smaller activity center located in the northeastern and southwestern part of Midtown Manhattan. The activity centers emerges around noon, then become prominent during nightfall, and in the late evening, only the two major activity centers remain while the smaller activity centers fade out. For the recreation and shopping activity, the activity centers remains stationary, only the intensity level changed at different time of the day. Thus the popular places of these activity categories seems have consistent ability of attracting visitors compare with other less popular places.

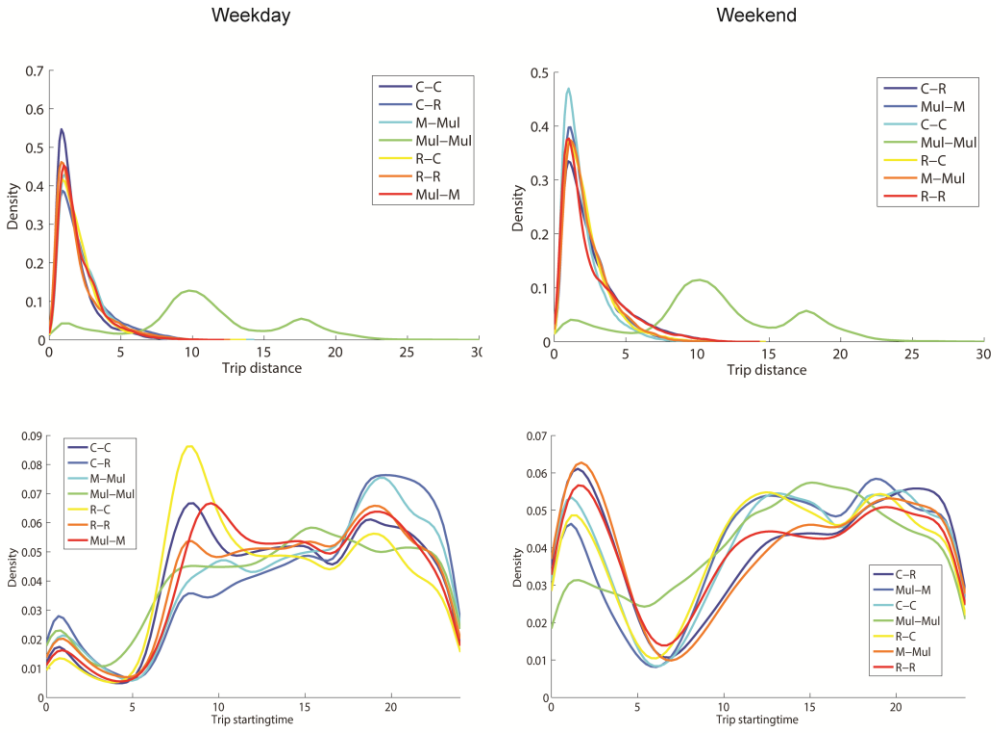


Figure 3.8 Distribution of Trip Distance and Trip Starting Time

Though taxi trips are mostly commercial and residential related, we observe that the Mul-Mul-L group is a very special type of taxi trips with unique characteristics. Based on the trip location distribution, these trips connect midtown Manhattan, LGA and JFK to the rest of NYC. While all other clusters have very short trip distance, the mean travel distance of the group reaches 11 miles. Two peaks are revealed from the distance distribution, which locate at 10 miles and 17 miles. The two points are matched with the travel distance from Manhattan to LGA and JFK respectively. As a result, the exclusive pattern is largely determined by the urban forms, as airports are usually far from the city center but with very high passenger volumes. The group of trip should be treated separately during urban studies as it is heavily biased from the general mobility pattern of taxi trips.

3.5 Human Mobility

Mobility pattern at individual level is often considered to be random, but the pattern over certain amount of population is barely random. Several studies using various data studying human mobility, including movement of an online game (Szell et al., 2012), the dispersal of bank notes (Brockmann et al., 2006) as well as trajectories from cellular data (Gonzalez et al., 2008) , have found highly regulated pattern in human movement. To be more specific, individual movement at an aggregated level is observed to follow a heavy-tailed plot under logarithmic scale. Same distribution is recognized for population of cities, the intensities of earthquakes, and the sizes of power outages and such kind of plots can be well approximated by power law distribution (Clauset et al. 2009).

Since human beings are the participants of taxi trips, we speculate if travel distance of taxi trips follows the same distribution. In order to reveal the underlying pattern, the travel distances are binned into bins by the width of 0.1 mile and the number of trips in each bin is counted and corresponding probability is calculated. The distribution of all the travel distance versus probability is plotted under logarithmic scale in Figure 3.9(a). Two different trends are revealed in the plot: travel distance is positively correlated with probability while distance smaller than 0.8 mile and negative correlation is observed for travel distance greater than 0.8 mile. Moreover, two inflections are captured at the distance around 10 miles and 20 miles. The pattern is consistent with previous discussion, as airport trips based at LGA and JFK give rise to the disturbances. For a better understanding of regular taxi trips, airport trips are removed and a refined plot is generated in Figure 3.9(b) where a smooth curve is presented.

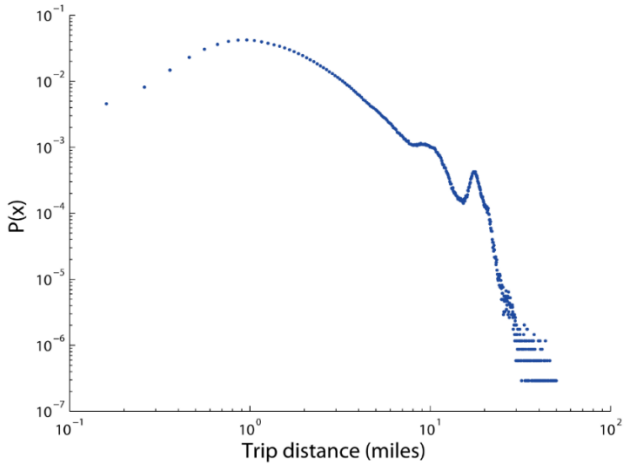
Trips with distance less than 0.8 mile take 16.89% of total trips. As very short trips within walking radius, these trips differ from the general pattern of taxi mobility on a decision making process of whether to take taxis. The first part of the trips can be approximated with distribution:

$$P(d) \propto d^{\beta} \tag{3.2}$$

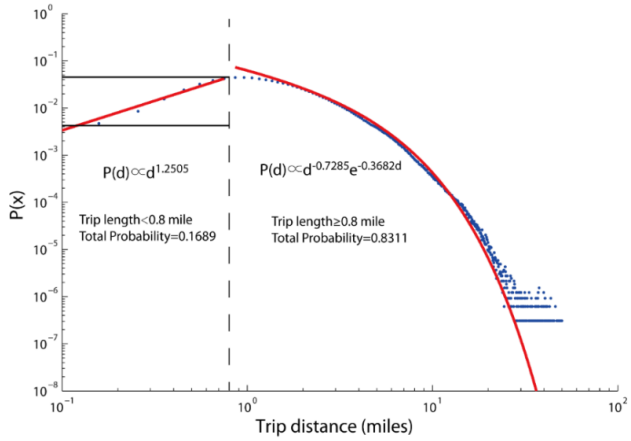
Where exponent β =1.2505.

The distribution resembles a power-law like distribution (straight line under logarithmic scale), however, the exponent takes a positive value. It indicates that for trips less than 0.8 mile, the probability of people making taxi trips increases with the travel distance and the relationship is about linear. The 0.8 mile threshold is an interesting

number. The average walking speed is around 3.2 miles per hour and is found to be positively correlated with Gross Domestic Product (GDP) per capita and population size (Levine and Norenzayan, 1999). Considering NYC as a city with large population and high GDP, the 0.8 mile is around a 10-minute walk. Consequently, the distribution implies a procedure of choice making. People are transferring from walking to taxicabs when travel distance increases, just as observed.



(a) Distance of all taxi trips



(b) Distance of all taxi trips excluding airport trips

Figure 3.9 Taxi Trip Distance and Human Mobility

The right part of taxi trips contributes 83.11% of total trips. Unsurprisingly, the scatted plot follows a heavy-tailed distribution which is observed in previous literatures. The difference lies in the inability to fit the plot by a straight line (power law distribution). However, the curve can be well approximated by a power law distribution with exponential cut-off (also known as truncated power law) which takes the form:

$$P(d) \propto d^{-\alpha} e^{-\lambda d} \tag{3.3}$$

With exponent α =0.7285 and λ =0.3682. Unlike the power law distribution of human movement reported, the taxi trip distance distribution has a faster probability decay in the tail part (the effect of the exponential cut-off term). This phenomenon should be explained from several aspects. Firstly, it indicates that the unique effects of urban environment on the distribution of taxi trip distance. Since the underlying size of urban area limits the distance of taxi trip, very long trip (e.g. > 30 miles) are less likely to happen, and the scale-free property of a typical power-law distribution fails. It is notable that as taxi trips are important component of urban human movement, the trip distance distribution reflects a unique perspective of human mobility. That is, the taxi mobility pattern reveals the hidden role of urban geographical boundaries in limiting urban human movement. Secondly, it reflects the self-selectivity behavior of people in the city. In a metropolitan such as NYC, a long distance taxi trip usually comes with expensive fare and long waiting time in traffic. The distribution implies that people view long taxi trips as not cost-effective and intend to avoid such trips. Thirdly, it supports the existence of competition among different transportation tools. Urban area generally has developed

transit system. For trip exceeding certain distance, subway or train is often a better substitution.

3.6 Summary

In this chapter, we exploit New York taxi trip data and comprehensively explore underlying patterns of urban taxi trips. A general framework is presented to make use of large scale data to understand to the pulse of a city. We first look at the general level of demand and find out the spatial and temporal patterns for the most popular places. The existence of unbalanced trip is inferred from the overall trip distribution and two typical types of unbalanced taxi trips are further analyzed. Given the spatial repeatability, we use the two-step clustering algorithm to figure out the intrinsic taxi trip classes. Patterns of different classes are discussed based on land use, travel distance and starting time distributions. In the end, the taxi mobility is found to partially follow a truncated power law distribution. The implications under the distribution are also explored. All the findings prove the large scale data as a promising tool in characterizing urban dynamics.

From the pattern recognized, we also delve into several meaningful insights. Unbalanced trips are common in taxi industry and should be carefully investigated. The existence will undoubtedly impair the equity of public transportation, lower the system efficiency, and degrade the level of service. Airport based taxi trip is identified as a very special part since it significantly differs from regular pattern. Land use has compelling impact on taxi trips and different types of taxi trips are able to uncover the structure of a city. Moreover, we discover that the mobility of taxi trips are restricted by the urban

geographical boundaries, human selectivity as well competition inside transportation system.

CHAPTER 4. MODELING INTRACITY MOVEMENT

4.1 Data Processing

The temporal and spatial repeatability is recognized from Chapter 3. Also, the mobility of taxi trips is found to follow a truncated power-law distribution. In this chapter, a comprehensive study is conducted to understand the reasons behind the repeatability pattern and the implications from the power-law distribution from an econometric point of view. Instead of checking origins and destinations separately, the data is processed to modeling urban human movement as a complete trip.

Considering the variation of trip over the week, the data is first divided into three time period: (1) morning peak from 7:00 am to 10:00 am, (2) off-peak hours from 12:00 pm to 3:00 pm, and (3) evening peak from 5:00 pm to 8:00 pm. Moreover, the weekday and weekend are modeled separately in light of the discrepancies revealed. For each piece of records, trip origin and destination are tagged with the corresponding zip code. As a result, the 168 ZCTAs generate 28,224 OD pairs and the dependent variable of the study is the amount of trips between each OD pair. The statistics of the dependent variable is given in Table 4.1.

Table 4.1 Summary Statistics for Dependent Variable (Trip Amount)

Time Interval	Mean	Std	Min	Max
Weekday Morning Peak	2.582	22.474	0	767
Weekday Off Peak	2.393	20.435	0	662
Weekday Evening Peak	2.578	20.79	0	626
Weekend Morning Peak	1.089	8.592	0	399
Weekend Off Peak	2.38	18.532	0	525
Weekend Evening Peak	1.095	8.32	0	265
Weekday Total	2.518	21.252	0	767
Weekend Total	1.521	12.749	0	525
Entire Week	2.233	19.215	0	767

4.2 Explanatory Variables

An open dataset (Geodatabase, Version: July 2013) from CUNY is referenced to obtain attributes of all ZCTAs. The database provides rich and useful information, i.e. 2010 census survey, American Community Survey, and facility locations in NYC. In total, 45 attributes, including demographics, education and income level, are extracted. Then, another open dataset PLUTO (Version: 13.2) from Department of City Planning are used to extract land use attributes of each ZCTA. In total, 8 attributes, including commercial area, residential area, official area, retail area, garage area, storage area, factory area and numbers of units, are obtained.

Among 53 attributes of ZCTAs, not all attributes are suitable to account for the taxicab trips. A selection of independent variables is developed according to previous results and from an intuitive point of view. Though there are few special studies

specifically related to taxi mobility, many studies are conducted in recent years for public transport services and daily mobility. Crotte (2008) concludes that the level of income influences travelers' attitudes on public transport. Matas (2004) suggests that the level of employment is a significant variable in explaining the subway demand. Liu et al. (2012) examine the temporal variations of both origin and destinations of taxi trips and their associations with different land use features. They also find that a typical residential area is a source area in morning time but a sink area in evening time and non-residential area has the opposite impact. Dargay and Hanly (2004) use the data from National Travel Survey of Great Britain for the years 1989-91 and 1999-2001 and find the importance of the land use factors considered on mode choice and car ownership. Jia and Harrington (2008) find that there is an increasing amount of recreational trips on Metrorail in Washington during evening.

Moreover, there are several important determinant factors in light of the characteristics of taxi trips. First, restricted by the pricing strategy of taxicabs, the distance between an OD pair should have significant impact on the amount of taxi trips. Similarly, if the commuters may have a longer commuting time, they should be less likely to take taxicabs. As a result, variables related to demographics, socioeconomics, land use, and travel attributes are selected as explanatory variables for further analysis. The detailed description of the selected variables is presented in Table 4.2

Table 4.2 Summary of Explanatory Variables

Variables	Description	Mean
Dist	Distance between OD pairs. (km)	2.389
Black	The black population density at a particular area. (km^{-2})	7.729
Jobs	The job density at a particular area. (km^{-2})	9.252
Unemp	The percentage of people having no jobs.	0.292
Highinc	1 if the area has average annual income greater than $115,000, 0 otherwise.	0.190
Land_mix	The level of diversity for the land use pattern. Ranging from 0 to 1, higher value means more diverse land use pattern.	0.512
T_commu	1 if the average commuting time is greater than 25 minutes, 0 otherwise.	0.464
Colleges	1 if the ZCTA has colleges, 0 otherwise.	0.042
Rec	1 if the number of recreational sites is greater than 3, 0 otherwise.	0.893

4.3 Methodology

4.3.1 Zero-Inflated Negative Binomial Model

Given the time interval, the intra-city movement is measured by the number of taxi trips travelling from one geographic unit (ZCTA) to the other. Since the numbers of taxi trips are non-negative integers, simply applying the ordinary least squares (OLS) regression may yield improper values (Washington et al., 2003). As a result, the model

should be carefully selected to account for the count nature of the data. Moreover, as presented in Table 1.b, the mean values of taxi trips for all time slots are much smaller than the variance ($E(y_{ij}) < VAR(y_{ij})$) which indicates the presence of over-dispersion. To address both characteristics of the data, the Negative Binomial model (NB) is thus implemented which takes the form:

$$\lambda_{ij} = EXP(\beta_i X_i + \beta_j X_j + \varepsilon_{ij}) \tag{4.1}$$

$$VAR(y_{ij}) = E(y_{ij})(1 + \alpha E(y_{ij})) = E(y_{ij}) + \alpha E(y_{ij})^2 \tag{4.2}$$

$$P(y_{ij}) = \frac{\Gamma(1/\alpha + y_{ij})}{\Gamma(1/\alpha) y_{ij}!} \left(\frac{1/\alpha}{1/\alpha + \lambda_{ij}}\right)^{1/\alpha} \left(\frac{\lambda_{ij}}{1/\alpha + \lambda_{ij}}\right)^{y_{ij}} \tag{4.3}$$

Where $EXP(\varepsilon_{ij})$ is the gamma-distribution error term with mean 1 and variance α^2, $\Gamma(.)$ represents gamma function; and α is the overdispersion parameter.

The equation 4.1 is the exponential function which accounts for the nature of non-negative integers. Equation 4.2 introduces the over-dispersion parameter, allowing the variance to differ from the mean. Another critical challenge in modeling taxi trip data is to address the existence of excessive zeroes. The intra-city movement can be viewed as products of urban economic behaviors. In regions with high population density and good economy such as midtown and lower Manhattan, people are more likely to make their trips by taxis. But for places with poor economy condition or far from the center area, people may feel costly to ride a taxi and the number of trips may drop drastically. As a result, there are many ZCTAs having no trips during a short time interval and this part of trips should be modeled separately.

The Zero-Inflated Count Data Model provides the flexibility to model the excessive zeroes for count data. In the model, there exist two qualitatively different states.

One state may result from simply failing to observe a taxi trip during the study period. The other state may be because of the inability ever to generate a taxi trip between a particular OD pair. The state may be raised from several aspects, such as the disutility of taxi trips or some places being against the willingness of taxi drivers. Simply applying the NB model which estimates the data as a single stage system may lead to erroneous inferences of overdispersion effect and potential determinant factors (Carson and Mannering, 2001). Alternatively, the Zero-Inflated Negative Binomial model (ZINB) is implemented to address the dual-state system as:

$$y_{ij} = 0 \text{ with probability } p_{ij} + (1 - p_{ij})\left(\frac{1/\alpha}{1/\alpha + \lambda_{ij}}\right)^{1/\alpha} \tag{4.4}$$

$$y_{ij} = y \text{ with probability } \frac{(1-p_{ij})\Gamma(1/\alpha + y)\mu_{ij}{}^{1/\alpha}(1-\mu_{ij})^{y}}{\Gamma(1/\alpha)y!} \tag{4.5}$$

$$\mu_{ij} = (1/\alpha)/(1/\alpha + \lambda_{ij}) \tag{4.6}$$

Where, *y* is the trip frequency from origin *i* to destination *j* during the given time interval.

In the part of NB model, the overdispersion parameter is used to examine the appropriateness against Poisson regression model. The null hypothesis for the over-dispersion is that the taxicab trips data are not significantly over-dispersed. The result provides the confidence level to reject the null hypothesis. For the zero state, previous studies suggest the implementation of zero-inflated models if more than 50% of the observations are zero (Miller, 2007). A more accepted metric, Vuong's value, is used to measure the model fitting of the ZINB. The null hypothesis for the Vuong's value is that there is no zero state and the NB is preferred than the ZINB.

4.3.2 Marginal Effect

Even for a significant variable, the effects may vary across different models. One may have a greater influence in one model than that in another model. Knowing the differences, more efficient strategies can be developed in traffic control and city planning. At here, marginal effect is introduced. In general, the term measures the effect that a unit change in an independent variable has on the response variable of interest. It is very popular in some disciplines (e.g. Economics) because they often provide a good approximation to the amount of change in dependent variables that will be produced by a unit change in independent variables. The term is helpful to learn the importance of an independent variable.

4.3.3 Multicollinearity Test

Before estimating models, there is still one important step to test the correlation among independent variables. Though highly correlated independent variables (multicollinearity) will not reduce the reliability of the model as a whole, it influences the computation regarding individual independent variables. As it may become infeasible to obtain the inverse of the matrix, or the inverse of the matrix may be calculated inaccurately. The regression model with multicollinearity can indicate how well the entire bundle of independent variables predicts dependent variable. However, the model cannot give valid information about any individual independent variable. The objective of the project is to study on the relationship between explanatory variables and taxicab trips. Removing or decreasing the effects of multicollinearity is beneficial for obtaining more accurate results.

We use the Variance Inflation Factor (VIF) to quantify the severity of multicollinearity among independent variables. Generally, if VIF is less than 10, the variables are seen free from severe multicollinearity issues and the estimated coefficients and marginal effects are considered as reliable.

4.4 Results

4.4.1 Model Estimations

As illustrated in Table 4.3, the VIF value of the independent variables are all smaller than 6. Moreover, the distance, unemployment rate, the land use mix, the commuting time, and the existence of colleges and recreational sites have the VIF value very close to 1. This implies that they merely suffer from the multicollinearity error. The highest value comes of the number of jobs which is 5.473. While it may have minor impact on the estimation results, however, the value is still in the safety range thus the results are still reliable.

Table 4.3 VIF Value for Explanatory Variables

Variables	VIF
Distance	1.021
Black	2.528
Number of Jobs	5.473
Unemployment Rate	1.275
Higher Annual Income	2.781
Land Use Mixture Index	1.326
Commuters' Mean Travel Time	1.409
Colleges	1.133
Recreational Sites	1.388

Table 4.4 presents the model comparison between ZINB and NB based on the estimated results. According to the log-likelihood at zero and at convergence, the ρ^2 takes the value from 0.25 to 0.29. Considering the large number of observations and complexity of the human movement, it provides a good proxy for understanding of the intracity movement in urban areas. The AIC value accounts for the goodness of model fitting while at the same time penalizing the model complexity. The huge reduction in AIC value indicates that the ZINB outperforms NB and provides better model quality. The θ value characterizes the existence of overdispersion. As suggested by the t-stat of θ, all six models are confident at 0.001 level to reject the null hypothesis which implies the data is overdispersed. The Vuong's value provides the statistical evidences that the data should be modeled as a dual-state system with more than 99.9% confidence. In

conclusion, the ZINB is deemed to be the appropriate model to analyze intracity movement compared with Poisson Regression and NB models.

Table 4.4 Model Comparison between ZINB and NB

	Weekday			Weekend		
Diagnostics	**MP**	**OP**	**EP**	**MP**	**OP**	**EP**
Zero-Inflated Negative Binomial						
Log-likelihood at Convergence	-69264.99	-70217.51	-71220.27	-25611.54	-30445.73	-23350.00
Restricted Log-likelihood	-93210.58	-90973.33	-97510.87	-34751.64	-39528.99	-31137.05
AIC	138588	140489	142494.5	51277.07	60941.47	46737.87
theta	0.291	0.242	0.325	0.353	0.293	0.346
log theta	-1.235	-1.418	-1.123	-1.042	-1.228	-1.061
t-statistic	-79.67	-71.654	-77.128	-40.732	-46.959	-36.262
Vuong	36.23	10.042	43.451	23.587	11.303	9.583
Negative Binomial						
Restricted Log-likelihood	-73783.07	-71338.49	-76823.25	-27543.52	-31368.05	-24136.69
AIC	147596.1	142711	153676.5	55115.04	62770.1	48303.39

4.4.2 Non-Zero State

Table A.1 and Table A.2 present the model estimation results for trip generation ability at origins and trip attraction ability at destinations. There are 11 significant variables at the confidence level of 95% regardless of time, but the effects of each variable may vary over time. Among these 11 variables, the increase in travel distance, black population, being in high income group, and of longer commuting time will reduce the number of trip generations and attractions. The increase in number of jobs, colleges and land use mix are found to have positive effects on the number of taxi trips.

The longer distance and more travel time the trips are with, the fewer the taxicab trip frequency will be. Considering the relative expensive fares, taxicab trip with long distance and much travel time is not a better choice for passengers. Moreover, taxicab drivers may refuse the remote destinations as they have to run the risk of returning empty for a long trip. As a result, both reasons may reduce the taxicab trip frequency between two remote areas and decrease the attraction ability in the area with higher commuting time. The marginal effects of distance show that the value is lower at morning and evening peak hours compared with off-peak period. It implies the significance of taxicab as a commercial commuting tool and there is a rigid demand for taxi in rush hours, despite the high cost. On weekend, with the decrease in transit accessibility (e.g. reduced operation hours), some residents may have to transfer to taxicabs. The marginal effects of commuters' travel time at destination show that the drop in trip amount is more significant during peak time. Since areas with higher commuting time are far from city center and have a reduced activity intensity, it is therefore less attractive for taxicab drivers. For black people, they may be limited by their education level and may have relatively lower income level. Moreover, areas with more black population may have worse economy and more crimes. Hence, black people are less likely to afford the cost incurred by taxi trips and drivers may tend to avoid visiting these places. Areas with high annual income at both origins and destinations are observed to have fewer taxi trips. A possible explanation is that people with high income may purchase their own vehicle instead of taking taxicabs. The parameter has a weakened effect during weekday peak time, which is the time with the most traffic of the day. This may be because that rich people have more flexibility to adjust their departure time.

There are also three variables which contribute to the increase in taxi trip generation and attraction in NYC. The area with more job opportunities, likely full with commercial activities, may generate and attract more business trips. The taxicab is a better choice to complete these trips, confirmed by higher increases during off peak. During weekend evening peak, the area may be also with many people joining leisure activities and a huge demand on taxicabs emerges. The effects during weekend evening peak will be stronger. The existence of colleges are also significant generators and attractors. The free and open colleges make the area with more flows and more demand on taxicab trips on the whole day. Especially on weekend, without classes and more activities, the area with colleges will generate more taxicab trips. The land use mix characterize the diversity of different land uses in the area. In general, higher land use mix refers to places close to city center and rural areas usually have single or two land use types. The marginal effect shows that more taxi trips take place during weekday morning peak and weekend evening peak at places with higher land use mix. This conforms to the definition of land use mix and presents the trends on how people are moving in the city at particular time.

There are also variables only significant at specific time periods. The unemployment rate has negative effects on trip generations during weekday off peak and on trip attractions during morning peak and off peak. The area with higher unemployment rate may have few and inactive business activities, which results in fewer demand on taxicab trips. The area with more recreational sites are observed to generate more taxicab trips during off peak and evening peak and attract more taxicab trips only during morning peak. The recreational sites are also observed to attract even more taxi trips during

weekday morning peaks. While being insignificant on weekend morning, it is likely caused by delayed opening time. After participating activities, there will be significant demand on taxicabs departing from the recreational sites during off peak and evening peak.

4.4.3 Zero State

Table A.3 and Table A.4 present the model estimation on probability of people having no incentives to move around the city at particular time period. While the reasons for being in zero state is highly complex, the increase in trip distance, longer commuting time at origins, high unemployment rate at destinations, and high annual income at destinations may explain the implicit causalities of no taxi trip. Longer travel distance represents higher travel expenses. It also increases the probability of leaving Manhattan area and reaching rural areas. Therefore, people may less likely to choose relative expensive taxicabs when making long trips and some taxicab drivers may reject the trip to avoid empty trips. The longer commuting time at origins is not significant in the non-zero state, however, it is observed to be an important factor related to zero taxi trip. The variable represents places far from the city center. As a result, commuters may have less chance to find a taxicab. The marginal effects are especially stronger during weekday evening peak and during weekend morning peak. The unemployment rate and high annual income may fail to generate taxi trips. The reasons are similar to that in the non-zero state.

There are other variables that reduce the probability of having no taxi trips. The existence of colleges implies the existence of more young people and the potential of

generating more activities, which in return producing more taxi trips. The marginal effect suggests the likelihood of being in non-zero state is lower, probably due to the conflicts with academic activities. Intuitively, the marginal effect of the variable has its highest value during weekend.

Same as in the non-zero state, some variables are revealed to be significant only within specific time interval, including the number of jobs at origins, the high annual income and the land use mix at origins. More jobs will attract more commuters and generate more commercial trips. This will undoubtedly be helpful in decreasing the probability of attracting and generating no taxicab trips. During peak time, the effects are stronger due to high proportion of commuting trips. The high annual income is found to contribute to zero taxi trip during peak time, mainly caused by high car ownership. The high land use mix value at origins mainly refer to places within Manhattan, where more activities are agglomerated. The large amount of morning commuters in the morning and after business activities or other leisure activities in the evening eventually decreases the probability of attracting no taxicab trips. Considering the operation hours, the recreational sites mainly attract population during morning peak and off peak and few people will go there during evening peak. Thus, more recreational sites may help to avoid the zero state. The marginal effects will be stronger on weekend which can be understood from having more spare time.

4.4.4 Summary and Discussion

Given the information on travel-related variables (distance and commuters' travel time), demographic variable (black population), socio-economic variables (jobs, annual

income and unemployment rate) and built environment variables (land use mixture, colleges, and recreational sites), it is obvious that the work can specify the intra-city movement using taxicab trips. A further increase in trip distance and travel time will decrease taxicab trip frequency and increase the probability of having no taxicab trips. However, the effects are weakened during peak time and on weekend, considering many fixed commuting trips during peak time and leisure trips on weekend. The active socio-economic activities, i.e. more number of jobs and low unemployment rate, will attract and generate more commuting trips and business trips, and be helpful in increasing taxicab trips and reduce the risks of no taxicab trips, especially during peak time. The areas, lived with many rich population or black population, fail to generate and attract more taxicab trips and will increase the probability of having no taxicab trips, especially on weekdays, due to high car ownerships and relative lower income, respectively. The high mixed land use, with colleges and with more recreational sites are helpful in attracting and generating taxicab trips and decreasing the probability of having no taxicab trips. Especially on weekend, open colleges and recreational sites will attract many visitors and generate high demand on taxicab trips after activities.

The above results may provide a good reference for taxicab agencies and planning departments. For planning departments, the importance of land use mixture and built environments in taxicab trips and potential interactions are confirmed. Highly mixed land use, introducing more colleges and recreational sites, are all related to more taxicab trips. Through controlling the educational and recreational units and mixing land use, a more appropriate number of trips can be generated and attracted, which can ensure a satisfying level of service with existing facilities. For taxicab agencies, a more efficient cruising

strategy can be developed for drivers to avoid high vacant rate. During morning peak and evening peak, taxicab drivers can search passengers around areas with highly mixed land use, more jobs and some educational and recreational units. Meanwhile, some areas with high annual income, more commuter travel time and more black population, should be avoided. Finally, the taxicab agencies can also make a more sustainable expansion plan according to the characteristics of intra-city movement. With varying variables, they can predict the amount of intra-city movement in future and maintain a reasonable amount of taxicab medallions.

4.5 Conclusion

The paper focuses on characterizing intra-city movement using taxicab trips data recorded by GPS in NYC. The taxicab trips data for one week are aggregated at ZCTA level. Furthermore, the taxicab trips data are summarized in three time slots (MP, OP, EP) for each day to differentiate the temporal effects on taxicab trips. Various attributes of ZCTAs (including demographic, socio-economic and land use etc.) are collected as explanatory variables. In light of the count data nature of the dependent variable, Negative Binomial model and Zero-Inflated Negative Binomial model are estimated and the results are compared. ZINB outperforms NB considering the modeling fitting, the ability to address overdispersion and the excessive zeroes.

Though the Zero-Inflated Negative Binomial model can specify the population movement well, there are still some limitations. First, different attributes may have different effects across locations. If introducing random parameters, the results will be more useful. Second, the used 3-hour taxi trips dataset may eliminate the huge variance

during the 3 hours, which may introduce errors in final results. Finally, to learn the more detailed intra-city mobility, more various human trips data should be collected. Taxicab trips are mostly made by people with good income level and with short distance. If transit trips data are available and included, the results about intra-city mobility will be more accurate and meaningful.

CHAPTER 5. TAXICAB RIDERSHIP ANALYSIS

5.1 Taxicab Ridership

Despite an extensive study in transit ridership, few of them look into the ridership of taxicabs. The importance of forecasting taxi ridership can be understood from two aspects. Firstly, stakeholder wants to know how changes in land use or network structure may affect the demand of taxi trips. Secondly, as for taxi commissions, the most important issue to address is how to control the number of taxi medallions. Improper marketing strategy leads to either too many taxis on the road which makes it hard for drivers to make profit, or insufficient supply in which passengers can hardly find a vacant taxicab. For big cities like NYC, the taxi market has a long history and is in a good health. As a result, understanding the relationship between current taxi ridership and the structure of the city is very helpful to not only future planning but also a good guidance for other cities that is planning and developing their taxi systems. In this chapter, two models are estimated for weekday trips and weekend trips. The dependent variable is the taxi demand which is considered as the total number of trip origins in each ZCTA. The independent variables are selected with concentration on built environment and social functionalities.

5.2 Data Preparation

5.2.1 Study Area

The study area covers the majority part of the NYC, including Manhattan, Bronx, Brooklyn, and Queens. The Staten Island is excluded due to very few trips recorded. The ZIP Code Tabulation Areas (ZCTAs) are used as the geographical representation of the NYC. It provides a reasonable scale to understand the spatial pattern of the urban taxi ridership and can be easily correlated with the demographics and socioeconomics. The ZCTAs in small size are merged into the peripheral polygon to avoid the issue of the spatial misrepresentation (Grubesic and Matisziw, 2006). The final shape file of the study area contains 168 ZCTAs and all pick-up locations are aggregated into the corresponding ZCTA.

5.2.2 Dependent Variable

The taxi ridership is obtained by aggregating pick-up locations into the corresponding ZCTA. The weekday and weekend ridership are calculated as the average daily value. An underlying fact of urban taxi ridership is that most trips are concentrated at particular areas, for example, city center or CBDs. As a consequence, the histogram of the ridership distribution is highly skewed. To improve the interpretability and satisfy the normal distribution, the log transformation is applied to the ridership data and the result is presented in Figure 5.1.

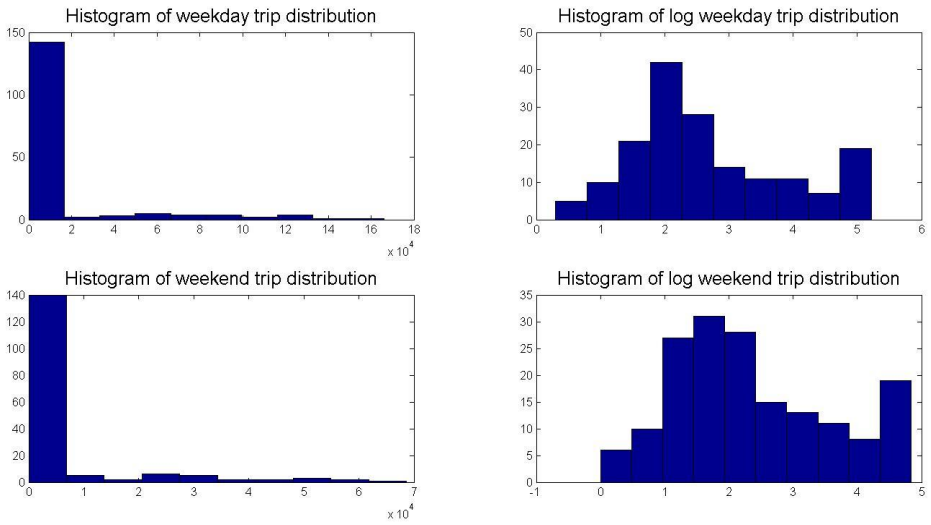


Figure 5.1 Data Transformation for Taxi Ridership

5.2.3 Explanatory Variables

The NYC Geodatabase is used to obtain the explanatory variables along with the ZCTA shape file and the road network shape file. The database includes various data geographies such as the 2010 census survey, American Community Survey (ACS) and a wide collection of facility locations (colleges, hospitals, metro stations etc.) in NYC.

Table 5.1 Candidate list of explanatory variables

Category	Name	Description
Demographic	Hispanic Population	Total number of Hispanic people in each ZCTA
	BS Population	Total number of people with a bachelor degree or higher in each ZCTA
	Employment	Total number of people under employment in each ZCTA
	Median income	The median income among all population in each ZCTA
Land Use	Commercial Area	The area used for commercial purposes in each ZCTA (m)
	Residential Area	The area used for residential purposes in each ZCTA (m)
	Road density	The length of road ways per square meters in each ZCTA
	Bike lane density	The length of bike lanes per square meters in each ZCTA
	Parking spaces	The number of parking spaces in each ZCTA
Travel Related	Subway accessibility	The accessibility subways in each ZCTA, higher value means easier to take the subway
	Bus accessibility	The accessibility to buses in each ZCTA, higher value means easier to take the bus

In the first step, a total of 11 explanatory variables are selected from three categories: demographics, land use and the accessibility to other public transport modes. Log-transformation is also applied to independent variables as closer linear relationships are observed under logarithm scale. The list of the 11 independent variables is given in

Table 5.1. For demographics, area population and employment are well recognized casual factors that influence the transit ridership (Taylor and Fink, 2003). Taylor et al. (2009) found that the median household income and percent of college students are significant factors to explain the ridership variations. Also, commercial and residential areas are two important factors related to the ridership (Sung and Oh, 2011). Parking supply is found to be the most significant factor in some studies (Morrall and Bolger, 1996; Chung, 1997). Besides, we also include the density of road network and bike lanes which are calculated as:

$$d_i = \frac{\sum_j L_{ij} n_{ij}}{A_i} \tag{5.1}$$

Where d_i is the i_{th} ZCTA, L_{ij} and n_{ij} are the length and number of lanes of link j in the i_{th} ZCTA and A_i denotes the area. The spatial distribution of the road density is plotted in Figure 5.2(a). The road density and bike lane density are hypothesized to be positively correlated with the ridership since they characterize the ease of access to a certain area.

In light of the special role of the taxicab in urban transit system, the third category is introduced to provide an understanding of the intrinsic relationship between taxis and other transport modes. There are 21 subway lines and 421 subway stations. To better calculate the accessibility to subways, the complexes of transfer stations are also taken into consideration which give 468 stations in total. Following the case study in Madrid (Gutierrez et al., 2011), the number of trips decreases with increasing walking distance and the relationship is approximately linear. The subway accessibility is calculated following:

$$SubA_i = \frac{1}{K}\sum_k \sum_j \frac{1}{d_{kj}^i} \tag{5.2}$$

Where d_{kj}^i represents the distance from node k in ZCTA i to subway station j, which is calculated using Haversine formula (Sinnott, 1984). Figure 5.2(b) shows the spatial pattern of the subway accessibility. For bus accessibility, a negative exponential relationship with respect to walking distance is revealed in existing literature (Zhao et al., 2003) and a minor modification is applied to swap from linear to negative exponential relationship on the distance calculation in equation (2).

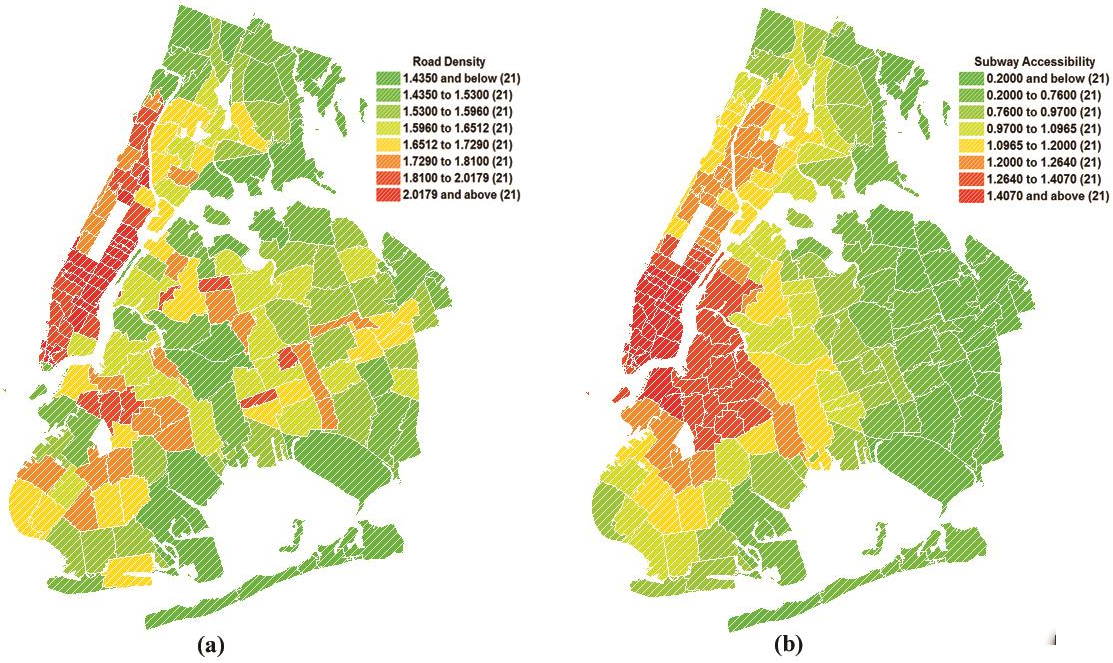


Figure 5.2 Spatial Distribution of (a) Road Density and (b) Subway Accessibility

5.3 Methodology

5.3.1 Multicollinearity

Multicollinearity is the problem where independent variables are correlated with each other. The appearance of the multicollinearity will not affect the accuracy of model estimation, however, it introduces bias in interpreting the significance and influence of a particular explanatory variable. While the objective of the study is to investigate the factors related to taxi ridership, it is of great importance to get rid of the potential multicollinearity among the variables. The multicollinearity is assessed following two steps. First, the Pearson product-moment correlation coefficients are calculated. Variables with coefficients greater than 0.7 are removed. Secondly, the variance inflation factor (VIF) is computed following the OLS analysis. The VIF is an indicator for the severity of the multicollinearity and the variable with VIF greater than 10 should be eliminated.

5.3.2 Spatial Autocorrelation

A critical drawback for the OLS is the assumption that the effect of the explanatory variable is fixed over the space. Due to the inherent functionality of the urban forms, however, the spatial patterns of both dependent and explanatory variables are stationary over the space. The spatial distribution patterns of the 11 variables are investigated by performing the global Moran's I test, which measures the spatial autocorrelation of a particular feature based on the location and numerical value simultaneously. The null hypothesis is that there is no spatial autocorrelation and the test statistic provides the confidence level to reject the null hypothesis.

5.3.3 Geographically Weighted Regression

The GWR model is an advanced tool to account for the spatial non-stationary over the space which allows for the coefficients varying locally. It can be viewed an extension of the multiple regression models by associating independent variables with geographical locations, which takes the following form:

$$y_i = a_{i0} + \sum_{k=1}^{n} a_{ik} x_{ik} + \epsilon_i \tag{5.3}$$

Where y_i is the dependent variable at location i and a_{ik} is the coefficient of x_{ik} at location i. Independent local models are estimated for each of the locations. The parameters are calibrated in the fashion that an observation will have greater impact on location i if the distance is closer. The impact is evaluated in a weighting scheme and is determined using the kernel function. In our case, the commonly-used Gaussian kernel function is implemented:

$$w_{ij} = \begin{cases} \exp\left[-0.5\left(\frac{d_{ij}}{b}\right)^2\right], d_{ij} < b \\ 0, otherwise \end{cases} \tag{5.4}$$

d_{ij} represents the distance between observation j and location i, which is calculated as the sphere distance in our study. The bandwidth b is used to exclude data point that exceeds the distance threshold. The GWR will gradually reduce to OLS as the bandwidth increases and will suffer the over-fitting problem if the bandwidth goes to zero. Considering that ZCTAs are denser near downtown areas and sparser at the peripheral, an adaptive bandwidth is use and the optimal bandwidth is determined by finding the

corresponding value that result in the minimum corrected Akaike information criterion (AICc). The results are carried out using the GWR 4.0 software (Nakaya et al., 2009).

5.4 Results

Table 5.2 presents the test result for the pairwise correlations between independent variables. It is observed that most of the correlation coefficients are below 0.7. However, the coefficients between the variables Hispanic population and people with bachelor degree (0.897), Hispanic population and employment (0.876), and employment and people with bachelor degree (0.812) all imply the existence of collinearity. As a result, no more than one of the three variables should be included in the model. Table 5.3 shows the results for the spatial autocorrelation. All variables are significant at 0.01 level, indicating the null hypothesis should be rejected. Moreover, since the z-score values are all positive, it implies that spatial distribution of variables are more likely to be spatially clustered. Therefore, it is highly desirable to use the GWR model to explore the spatial heterogeneity in the data.

The global models (OLS) are first calibrated to investigate the significant factors that influence the urban taxi ridership. Five independent variables are revealed to be closely related to the urban taxi ridership in both weekday and weekend model, including the population of bachelor degree, the median income level, the residential area, the road density and the subway accessibility. The detailed results of model estimation are presented in Table 5.4. According to the adjusted R^2, with only five explanatory variables, 61.47% and 60.18% of the variance can be explained for the weekday taxi ridership and weekend taxi ridership. As a result, the global models provide a good understanding of

the urban taxi ridership. The VIF value ranges from 1.27 to 1.71. Since all VIF values are close to 1, it indicates that the independent variables are well selected and the multicollinearity issue is avoided. For the t-statistics, the median income takes the value of -1.11 and -1.18 for weekday and weekend model respectively, suggesting that the variable is 73.11% and 76.11% confident to be different from 0. The other four variables are all statistically significant, with the BS population being significant at 0.05 level and the other three variables at 0.01 level. While some of the variables can be interpreted intuitively, such as the positive relationship between BS population and the taxi ridership, it is hard to explain effects of the median income and the subway accessibility from the global results.

Table 5.2 Pearson product-moment correlation coefficient for explanatory variables

	Hispanic Population	BS Population	Employment	Median income	Commercial Area	Residential Area	Road density	Bike lane density	Parking spaces	Subway access	Bus access
Hispanic Population	1										
BS Population	**0.897**	1									
Employment	**0.876**	**0.812**	1								
Median income	-0.079	0.239	-0.017	1							
Commercial Area	0.611	0.554	0.641	-0.249	1						
Residential Area	0.427	0.378	0.365	0.068	-0.390	1					
Road density	-0.256	-0.091	-0.353	0.335	-0.268	-0.141	1				
Bike lane density	-0.243	-0.086	-0.300	0.257	-0.164	-0.196	0.466	1			
Parking spaces	0.003	0.081	-0.040	-0.023	0.180	-0.258	0.411	0.432	1		
Subway access	-0.029	-0.006	-0.029	0.026	0.124	-0.258	0.531	0.507	0.517	1	
Bus access	0.092	0.147	0.116	0.056	0.213	-0.166	0.419	0.125	0.364	0.295	1

Table 5.3 Moran's I test result for candidate independent variables

Variables	**Moran's Index**	**Expected Index**	**z-score**	**p-value**
Hispanic population	0.406	-0.006	14.142	0.000
BS population	0.195	-0.006	6.873	0.000
Employment	0.479	-0.006	16.382	0.000
Median income	0.199	-0.006	10.213	0.000
Commercial area	0.109	-0.006	3.877	0.000
Residential area	0.560	-0.006	18.968	0.000
Road density	0.694	-0.006	23.512	0.000
Bike lane density	0.570	-0.006	19.365	0.000
Parking	0.390	-0.006	13.219	0.000
Subway accessibility	0.718	-0.006	24.425	0.000
Bus accessibility	0.256	-0.006	9.010	0.000

Table 5.4 Estimation results for global models (OLS)

Variable	Weekday		Weekend		VIF
	Coefficient	t-stat	Coefficient	t-stat	
Intercept	-0.520	-0.708	-0.481	-0.643	
BS population	0.221	2.143	0.214	2.040	1.306
Median income	-0.168	-1.054	-0.190	-1.178	1.289
Residential area	-2.393	-7.821	-2.221	-7.130	1.267
Road density	1.961	6.586	1.974	6.512	1.712
Subway accessibility	0.597	4.879	0.626	5.019	1.569
AIC		391.361		397.409	
AICc		392.0601		398.109	
R^2		0.629		0.616	
adjusted R^2		0.615		0.602	

The results of the GWR models are obtained using the same set of independent variables. Table 5.5 presents the model estimation for weekday GWR model and weekend GWR model.

Table 5.5 Estimations of the GWR models

Weekday						
Variable	Min	Max	Mean	Lower Quartile	Upper Quartile	DIFF
Intercept	-15.094	3.293	-6.325	-11.564	0.517	-187.017
BS population	-0.393	1.097	0.387	0.099	0.732	-104.247
Median income	-0.767	2.977	0.992	-0.143	1.870	-301.755
Residential area	-2.712	-0.832	-1.743	-2.146	-1.332	-13.453
Road density	0.042	2.061	1.172	0.861	1.547	-142.633
Subway accessibility	-0.317	4.524	1.148	0.545	1.576	-19.109
Weekend						
Variable	Min	Max	Mean	Lower Quartile	Upper Quartile	DIFF
Intercept	-15.802	3.189	-6.240	-11.462	0.517	-213.376
BS population	-0.324	1.044	0.379	0.106	0.750	-121.641
Median income	-0.775	3.112	0.935	-0.157	1.814	-327.283
Residential area	-2.468	-0.510	-1.518	-1.893	-1.162	-20.451
Road density	0.071	2.188	1.064	0.729	1.382	-154.409
Subway accessibility	0.011	4.945	1.421	0.585	1.946	-18.613

According to the median, the lower quartile and the upper quartile value, the estimated parameters for the BS population and the median income are stable over the week. The other three variables show a moderate disparity between the weekday and weekend model. As discussed before, the GWR model is used to account for the spatial non-stationary and the coefficients of independent variables vary from min to max. The DIFF value is an indicator of the spatial variability which evaluates the model performance between the global model and the GWR model. The negative DIFF values suggest that all five independent variables vary significantly over the space. As shown in

Table 5.6, the adjusted R^2 is 0.871 for weekday model and 0.875 for weekend model, which correspond to 0.256 and 0.273 increments in the amount of variation explained.

Table 5.6 Comparison between Global Model and GWR model

Diagnostics	Weekday		Weekend	
	Global	GWR	Global	GWR
AIC	391.361	214.644	397.409	209.700
AICc	392.061	226.872	398.109	221.927
R^2	0.629	0.899	0.616	0.903
Adjusted R^2	0.615	0.871	0.602	0.875
Residual sum of squares	92.970	25.112	96.379	24.383
F-value		11.353		12.405

Moreover, the reduction on the AICc value indicates that GWR provides much better goodness of fit given the same set of data comparatively. Great improvements are also achieved for the residual sum of squares and changes of the residual value are plotted in Figure 5.3. It is revealed that the spatial residual patterns are quite similar between weekday and weekend models. The residual value is reduced for most of the ZCTAs. Specifically, central Manhattan, lower Manhattan and the JFK airport receive the most notable reductions. More importantly, the global Moran's I index for the residual was 0.337 and 0.338 in the weekday and weekend global model. The new Moran's I values reduce to 0.161 and 0.133 respectively, which reflect the mitigation for the spatial autocorrelation in the data.

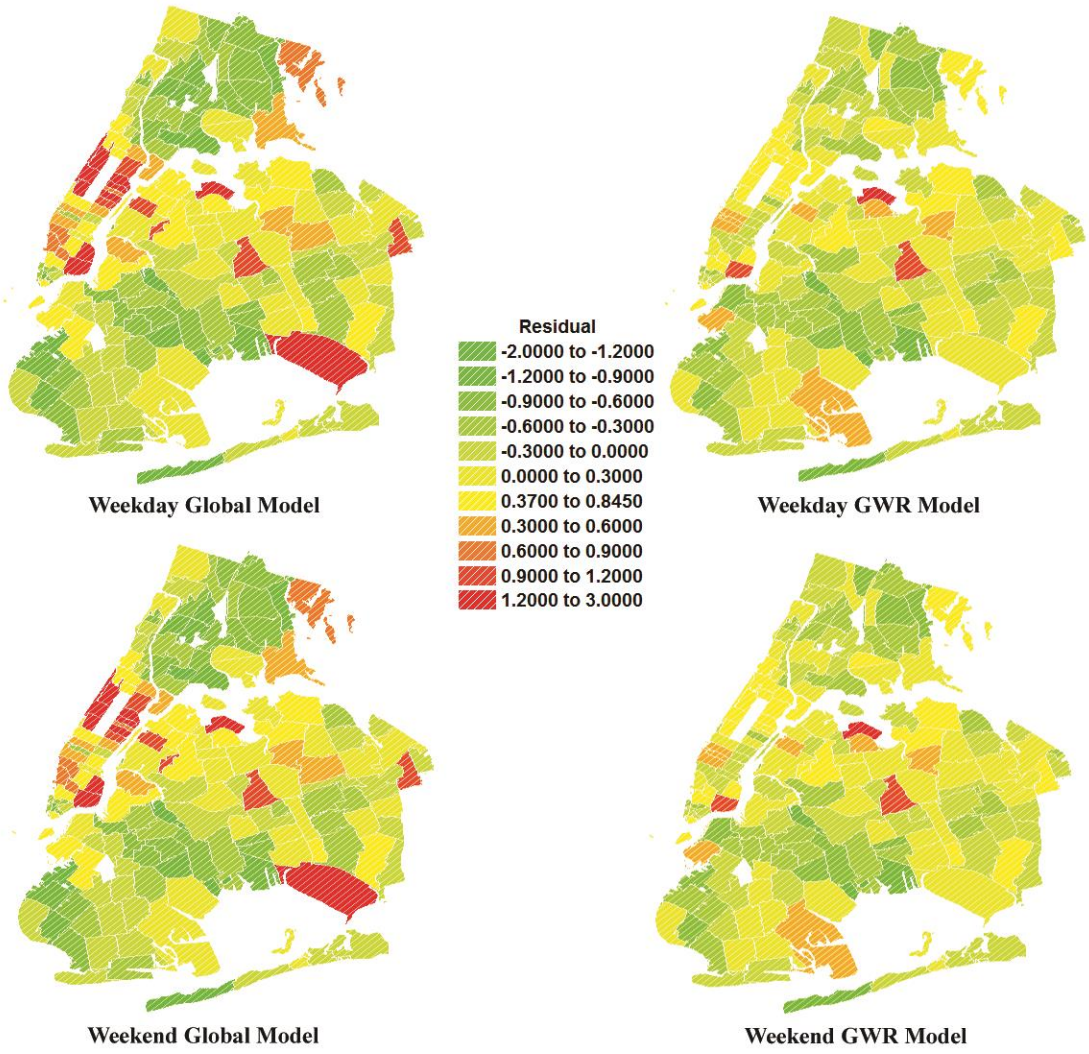


Figure 5.3 Spatial Distribution of the Residual

5.5 Discussion

It is evident that the GWR model outperforms the global model in explanatory power, goodness of model fit and accounting for the spatial non-stationary. Moreover, the GWR model also provides an in-depth understanding of how the variables are varying locally.

In the global model, the parameter specification for the weekday model is consistent to that of the weekend model. The median income level and the residential area are observed to be negative correlated with the number of taxi ridership. The increments in population with bachelor degree or higher, the road density and the subway accessibility have positive contribution to the taxi ridership. However, the relationship between the median income level and the ridership is counterintuitive and not statistically significant (0.05 level). The results of the GWR model suggest that the influence of the median income level may be either positive or negative, which depends heavily on the geographical location. Figure 5.4 visualizes the spatial distribution of the coefficients and the t-stats for the median income level during weekdays.

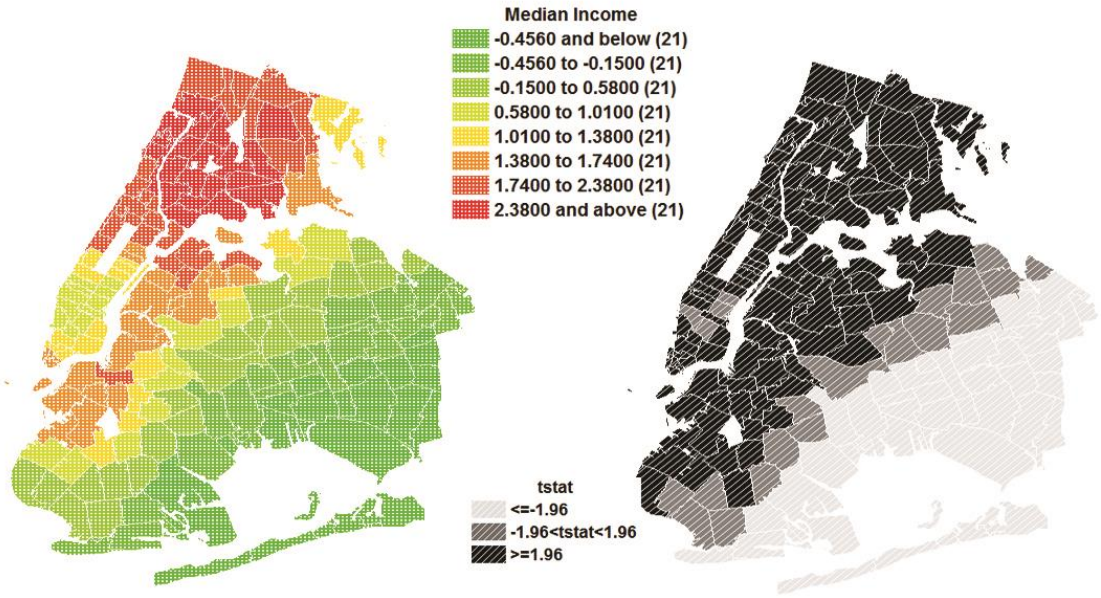


Figure 5.4 Spatial distribution for the coefficients and t-stats of the median income level

It can be seen that the coefficients are negative only at Northeast Brooklyn and East Queens, which indicates the functionality of the urban form: there are more open spaces

and less congestions in the area far from the central of the city. Moreover, it is costly to ride a taxi. As a result, people with higher income may own their private vehicles thus reducing the chance of hailing a taxicab. On the contrary, in places near city center such as the congested Manhattan, the coefficient turns to be positive due to the increment in the utility of riding taxis. The coefficient has its highest value around the Bronx area. A possible explanation is that the Bronx has a large population density and residential area, but also a high percentage rate of people under poverty line. While private vehicles may not be affordable, the higher income gives rise to the increasing taxi ridership. The distribution of t-stat also proves that the influence of the median income level is associated with the urban structure, since it is statistically significant for places close to or far from the central area and insignificant in the middle.

The same pattern is observed for the BS population that there are the variable has a negative impact on the taxi ridership in approximately 20% of the ZCTAs. As shown in Figure 5.5, more BS population brings more taxi ridership in the western part of NYC but decreases the ridership in the eastern part, especially Bronx. Moreover, the t-stat suggests that the only parameters associated with western part are statistically significant. Therefore, the estimations in the global model is biased. People who received education level of bachelor or higher are more likely to find better jobs with higher payment. In return, these jobs are mostly located in the central area of the city and these people have the opportunity to leave closer to the city center. The comparatively higher salary and the place where they work and live imply that they may use taxis more often, which corresponds to the spatial distribution of the coefficients.

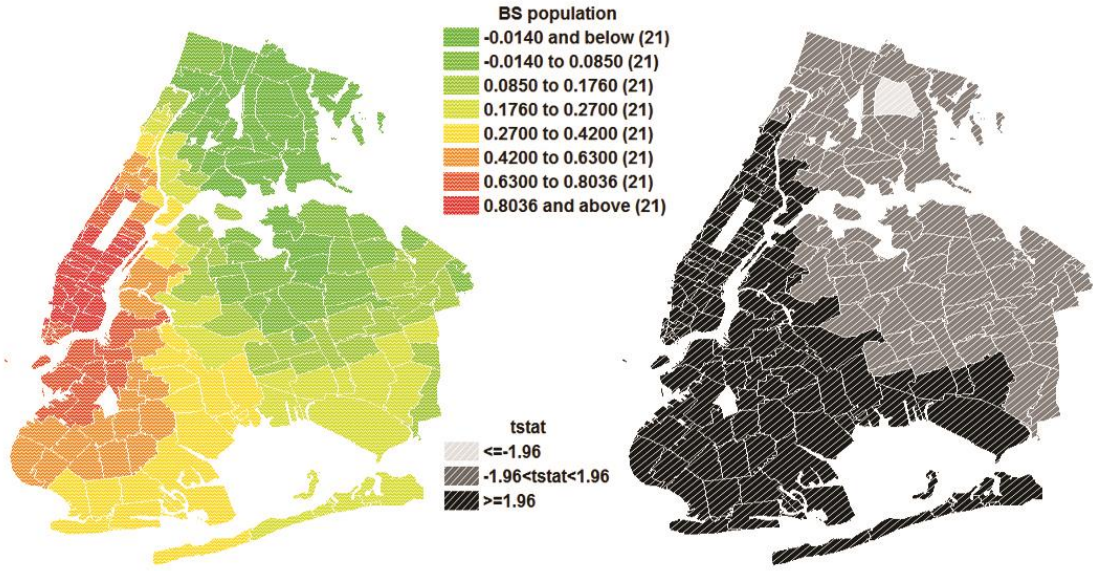


Figure 5.5 Spatial distribution for the coefficients and t-stats of the BS population

The rest of the three independent variables are observed to be statistically significant in the entire study area. For parameters of the residential floor area, the negative coefficients should be understood from the inherent characteristics of NYC taxi trips and may not be the same for other cities. Qian et.al (2013) has revealed that around 20% of the trip origins are associated with residential land use and most of the trips are commercial oriented. An increased number of residential areas implies the decrease in other types of land use which contribute mostly to taxi ridership thus leading to fewer trips. From the urban geography point of view, there are usually more residential areas as the distance from the city center increases. As mentioned earlier, the increasing distance also lowers the utility of taking taxis and the results in Figure 5.6(a) where the coefficients being smaller outside the city center conform the statement.

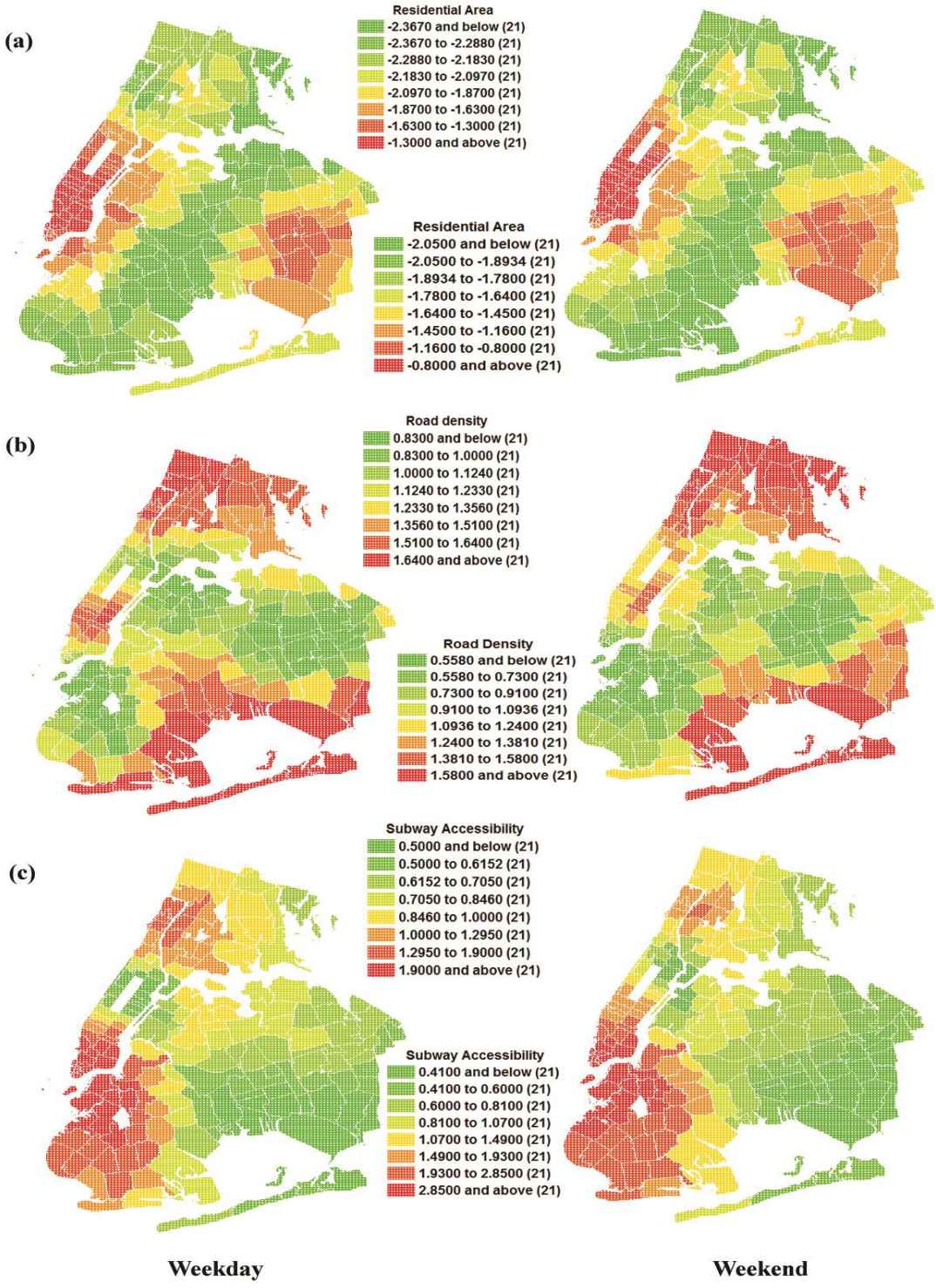


Figure 5.6 Spatial Distribution for variables statistically significant over the space (a) Residential Area, (b) Road Density and (c) Subway Accessibility

A special case is observed towards the most eastern corner in Queens probably due to the existence of the JFK airport. Though the general distribution is similar between weekday and weekend, the greater coefficient values suggest that residential areas generate more taxi trips during the weekend. Higher road density usually comes with higher population density and more accessibility for vehicles. Therefore, increasing in the road density will generate more taxi trips. The Figure 5.6(b) also indicates that the influence of the road density at the center and in the peripheral area of a city is comparatively higher than other places and is minor during the weekend. Finally, the parameter estimations for subway accessibility suggest that more taxi trips occur in the places with more exposure to subways. Due to the special functionality and pricing scheme, the parameters show no ridership trade-off between the two public transport methods. Instead, following Figure 5.2(b) and Figure 5.6(c), the spatial pattern of the coefficients comply with that of the accessibility value and even more taxi trips are generated during the weekend. The phenomenon can be explained from two aspects. First, places with subway stations indicate the existence of high passenger volume and are frequently visited, which at the same will attract and produce more taxi trips. Secondly, the taxicab may be widely used as a connection transport tool between initial locations/last destinations and subway stations.

5.6 Final Remarks

The taxi market in large cities is expanding immensely with the fast urbanization process, however, very limited efforts are made to understand the determinants related to taxi ridership. While there is an urgent need for an efficient tool to help framing better

policies, traditional approaches for ridership analysis such as multiple regressions may not fully characterize the complexity of the urban form. In this study, we present an efficient and accurate tool to tailor the need from the rapid development of taxi industry.

All estimations and findings suggest that taxi ridership analysis with GWR model may serve as a powerful method to predict the passenger demand for taxis. As demonstrated in Figure 5.3, the amount of residues for GWR model receives comprehensive improvement compared with the global model. Specifically, Manhattan area and two airports (LGA and JFK) has the greatest reduction residual value. This may help agencies to better estimate the new passenger demand for taxis when constructing new facilities such as a shopping mall. Since not only the level of demographics and socioeconomics of a particular area matters, the functionality and the geographical location of the place are also very crucial when determining the number of ridership. By addressing spatial non-stationary, the GWR model takes the functionality of urban form into consideration and therefore produces more precise and realistic estimations.

Moreover, the explanatory power of GWR model helps to recognize the impact of determinants correctly. Indicated by results from global models, some coefficient estimations are counterintuitive due to the inappropriate assumption when dealing with urban problems. As a matter of fact, the effects of ridership determinants are closely associated with the urban from and geographical boundary. The visualization helps to understand the heterogeneity of determinants and to better interpret the spatial variation of variable coefficients.

As a consequence, the prediction power together with the explanatory power of GWR model will undoubtedly contribute to the sustainable and efficient development of

the taxi industry. By estimating future ridership, it helps to regulate the overall number of license according to the predicted passenger volume. Moreover, the detailed understanding for the local variation of taxi demand helps to better distribute the available taxi resources and improves system efficiency. Last but not least, since the NYC taxicab is the benchmark in the industry, results in our study may provide beneficial insights for the development of new taxi market and serve as an exemplar for existing taxi market to enhance the current level of system performance.

CHAPTER 6. CONCLUSION

6.1 Summary

With the development of pervasive computing technology, the availability of large scale data brings unprecedented opportunities for researchers to explore several untouched areas. Especially in the booming urbanization process, the intensive human activities can be captured by various data sources such as the location of mobile phone users, the posts from social media websites, as well as the trip information from taxicabs. The big data carries invaluable information which may greatly help to tackle big problems in large cities. But they are not ready for use. There is an urgent need to develop appropriate tools to decode the data and understand the data.

In this thesis, we present a comprehensive study on utilizing NYC taxi trip data to characterize urban travel patterns. We first process the data to remove inconsistencies and extract the portion that matches our research needs. The work starts by recognizing urban patterns, trying to understand what the dynamics of urban activities is and exploring inherent homogeneity among various trips. We also reveal the uniform pattern of human mobility from taxi movements. Moreover, the geographical database of NYC is combined in our study to pursue further knowledge of urban travel demand. Since the taxi trip data captures very detailed spatiotemporal attributes, we use econometric and statistical tools to investigate determinant factors which influence the intracity population

movement by taxicabs and the spatial distribution of taxi ridership. The findings indicate the significant impact of urban form, land use functionality and geographical boundary on human mobility pattern and travel demand in urban areas.

The results of the study will undoubtedly contribute to urban planning and policy making. It offers beneficial insights for framing regulations for the taxi industry. More importantly, it builds up a framework to utilize the big data and explore the value of the big data in urban analytics.

6.2 Limitations and Future Work

The limitation of the work is mainly from the data. The geographical information contained in the dataset is only latitude and longitude coordinates for trip pick-ups and drop-offs. Due to privacy concerns, we are unable access to the full trajectory of taxi trips. This restricts our work from studying more detailed topics such as monitoring the congestion level of urban network, estimating the emission in urban areas and build recommendation systems for taxi drivers and passengers.

Another limitation is that we only focus on NYC alone to examine the urban travel patterns. While human travel behavior is revealed to have close connection with the urban structure, the result will be more convincing if we are able to conduct more case studies in other cities such as Hong Kong and Tokyo, which may help to clarify the detailed influence of urban forms on human activities.

6.3 Future Work

As extensions of the study, the following ideas may be considered to remedy defects of our work and further realize the value of the taxi data:

- Constructing a product using the ideas and methodologies in the study, such as a decision-making tool for planning agencies and administrators of taxi industries;
- Land use inference from taxi trip. In our study we reveal the homogeneity of taxi trips with respect to land uses and trip attributes. Therefore, it is meaningful to predict the land use pattern of urban areas from the spatiotemporal characteristics of taxi movements.
- Taxi recommendation system. From the hot spots analysis, trip distribution and trip classification, we obtain valuable information on how the taxi trips are distributed both spatially and temporally. Further, the econometric model helps us to understand the causalities and determinant factors that will generate and attract taxi trips. Based on the results, there is huge potential to provide both taxi drivers and passengers the useful information thus increasing the level of service for the taxi industry.
- The combination use of more data sources. We present the work of combing taxi trip data and geographical database to investigate the determinants influencing travel demand. There are lots of potential for other combinations, such as the mix use of social media data and taxi trip data for a complete study of urban human activities.

- More comprehensive study of urban travel patterns. We only use NYC taxi data from 2009. With the availability of the recent taxi trip data, it is important to explore the change of urban travel patterns. Also, it is beneficial to study the functionality of urban structure in different cities if the data from multiple cities is accessible.

REFERENCES

Agthe, D.E. & Billings, R.B., 1978. The impact of gasoline prices on urban bus ridership. The Annals of regional science, 12(1), pp.90–96.

Anderson, J.E. & Van Wincoop, E., 2004. Trade costs,

Andrienko, Y. & Guriev, S., 2004. Determinants of interregional mobility in Russia. Economics of transition, 12(1), pp.1–27.

Batty, M. & Xie, Y., 1994. From cells to cities. Environment and planning B, 21, p.s31.

Batty, M., Xie, Y. & Sun, Z., 1999. Modeling urban dynamics through GIS-based cellular automata. Computers, environment and urban systems, 23(3), pp.205–233.

Ben-Akiva, M.E. & Lerman, S.R., 1985. Discrete choice analysis: theory and application to travel demand, MIT press.

Bikker, J.A., 1992. An International Trade Flow Model With Zero Observations: an Extension of the Tobit Model.

Boyle, P.J., Halfacree, K. & Robinson, V., 1998. Exploring contemporary migration, Longman London.

Brockmann, D., Hufnagel, L. & Geisel, T., 2006. The scaling laws of human travel. Nature, 439(7075), pp.462–465.

Burger, M., Van Oort, F. & Linders, G.-J., 2009. On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation. Spatial Economic Analysis, 4(2), pp.167–190.

Calabrese, F. et al., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies, 26, pp.301–313.

Carson, J., & Mannering, F. (2001). The effect of ice warning signs on ice-accident frequencies and severities. Accident Analysis & Prevention, 33(1), pp. 99-109.

Cervero, R. & Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity, and design. Transportation Research Part D: Transport and Environment, 2(3), pp.199–219.

Chang, H. et al., 2008. iTaxi: Context-aware taxi demand hotspots prediction using ontology and data mining approaches. Proceedings of the 13th

Chiu, T. et al., 2001. A robust and scalable clustering algorithm for mixed type attributes in large database environment. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’01, pp.263–268.

Chow, L.-F. et al., 2006. Transit ridership model based on geographically weighted regression. Transportation Research Record: Journal of the Transportation Research Board, 1972(1), pp.105–114.

Chung, Kyusuk. "Estimating the Effects of Employment, Development Level and Parking Availability on CTA Rapid Transit Ridership: From 1976 to 1995 in Chicago." In proceedings. 1997.

Clauset, A., Shalizi, C.R. & Newman, M.E.J., 2009. Power-law distributions in empirical data. SIAM review, 51(4), pp.661–703.

Cristaldi, F., 2005. Commuting and gender in Italy: a methodological issue. The Professional Geographer, 57(2), pp.268–284.

Crôtte, Amado. Estimation of transport related demand elasticity in Mexico City: an application to road user charging. Master Thesis, Center for Transport Studies, Department of Civil and Environmental Engineering, Imperial College, United Kingdom. (2008).

Curry, L., 1972. A spatial analysis of gravity flows. Regional Studies, 6(2), pp.131–147.

Dargay, Joyce, Mark Hanly. Land use and mobility. World conference on Transport Research, Istanbul, Turkey. (2004)

De la Mata, T. & Llano-Verduras, C., 2012. Spatial pattern and domestic tourism: An econometric analysis using inter-regional monetary flows by type of journey*. Papers in Regional Science, 91(2), pp.437–470.

Deng, M. & Athanasopoulos, G., 2011. Modelling Australian domestic and international inbound travel: a spatial--temporal approach. Tourism Management, 32(5), pp.1075–1084.

González, M.C., Hidalgo, C. a & Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature, 453(7196), pp.779–782.

Grubesic, Tony H., and Timothy C. Matisziw. "On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data." International journal of health geographics 5, no. 1 (2006), pp. 58

Gutiérrez, J., Cardozo, O.D. & García-Palomares, J.C., 2011. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. Journal of Transport Geography, 19(6), pp.1081–1092.

Harris, B., 1985. Urban simulation models in regional science. Journal of Regional Science, 25(4), pp.545–567.

Hasan, S., Zhan, X. & Ukkusuri, S. V, 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing. p. 6.

Haynes, K.E. & Fotheringham, A.S., 1984. Gravity and spatial interaction models, Sage publications Beverly Hills.

Jia, Wendy, and Tom Harrington. "Metrorail Ridership Trends in the Washington Metropolitan Region." In 2008 American Public Transportation Association (APTA) Rail Conference. (2008).

Jiang, B., Yin, J. & Zhao, S., 2009. Characterizing the human mobility pattern in a large street network. Physical Review E, 80(2), p.21136.

Kanafani, A., 1983. Transportation demand analysis.

Karemera, D., Oguledo, V.I. & Davis, B., 2000. A gravity model analysis of international migration to North America. Applied Economics, 32(13), pp.1745–1755.

Keijer, M.J.N. & Rietveld, P., 2000. How do people get to the railway station? The Dutch experience. Transportation Planning and Technology, 23(3), pp.215–235.

Lambert, D., 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), pp.1–14.

León-Ledesma, M. & Piracha, M., 2004. International migration and the role of remittances in Eastern Europe. International Migration, 42(4), pp.65–83.

LeSage, J.P., Fischer, M.M. & Scherngell, T., 2007. Knowledge spillovers across Europe: Evidence from a Poisson spatial interaction model with spatial effects*. Papers in Regional Science, 86(3), pp.393–421.

Levine, R. V & Norenzayan, A., 1999. The pace of life in 31 countries. Journal of cross-cultural psychology, 30(2), pp.178–205.

Li, B. et al., 2011. Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset. In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on. pp. 63–68.

Liang, X. et al., 2012. The scaling of human mobility by taxis is exponential. Physica A: Statistical Mechanics and its Applications, 391(5), pp.2135–2144.

Linders, G.-J.M. & De Groot, H.L.F., 2006. Estimation of the gravity equation in the presence of zero flows,

Liu, Y. et al., 2012. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. Landscape and Urban Planning, 106(1), pp.73–87.

Matas, Anna. "Demand and revenue implications of an integrated public transport policy: The case of Madrid." Transport Reviews 24, no. 2 (2004), pp. 195-217.

Mattson, J.W., 2008. Effects of rising gas prices on bus ridership for small urban and rural transit systems, Upper Great Plains Transportation Institute, North Dakota State University.

McNally, M.G., 2008. The four step model.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis & Prevention, 26(4), pp.471–482.

Miller, J. M. (2007). Comparing Poisson, Hurdle, and ZIP model fit under varying degrees of skew and zero-inflation (Doctoral dissertation, University of Florida).

Morrall, John, and Dan Bolger. "The relationship between downtown parking supply and transit use." (1996).

Muller, P., 1995. Transportation and urban form. The geography of urban transportation.

Murray, A.T. et al., 1998. Public transportation access. Transportation Research Part D: Transport and Environment, 3(5), pp.319–328.

Nakaya, T., 2001. Local spatial interaction modelling based on the geographically weighted regression approach. GeoJournal, 53(4), pp.347–358.

NYCTLC, 2012. New York City Taxi and Limousine Commission 2012 Annual Report.

Pan, G. et al., 2013. Land-Use Classification Using Taxi GPS Traces. IEEE Transactions on Intelligent Transportation Systems, 14(1), pp.113–123. Available at:

Peng, C. et al., 2012. Collective human mobility pattern from taxi trips in urban area. PloS one, 7(4), p.e34487.

Phithakkitnukoon, S., Veloso, M. & Bento, C., 2010. Taxi-aware map: identifying and predicting vacant taxis in the city. Ambient …, pp.86–95. Available at:

Pucher, J., Hendrickson, C. & McNeil, S., 1981. Socio-economic characteristics of transit riders: some recent evidence. Traffic Quarterly, 35(3).

Rae, A., 2009. From spatial interaction data to spatial interaction information? Geovisualisation and spatial structures of migration from the 2001 UK census. Computers, Environment and Urban Systems, 33(3), pp.161–178.

Ratti, C. et al., 2006. Mobile Landscapes: using location data from cell phones for urban analysis. Environment and Planning B: Planning and Design, 33(5), pp.727–748.

Reades, Jonathan, et al., 2006. "Cellular census: Explorations in urban data collection." Pervasive Computing, IEEE 6.3 (2007): 30-38.

Shankar, V., Milton, J. & Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis & Prevention, 29(6), pp.829–837.

Silva, J.M.C.S. & Tenreyro, S., 2006. The log of gravity. The Review of Economics and statistics, 88(4), pp.641–658.

Sinnott, Roger W. "Virtues of the Haversine." Sky and telescope 68 (1984), pp. 158.

Song, Y. et al., 2012. Industrial agglomeration and transport accessibility in metropolitan Seoul. Journal of geographical systems, 14(3), pp.299–318.

SPSS, I.N.C., 2001. The SPSS TwoStep cluster component: A scalable component to segment your customers more effectively."

Sun, L., Chen, C. & Zhang, D., 2013. Understanding Urban Dynamics from Taxi GPS Traces. Creating Personal, Social, and Urban Awareness through Pervasive Computing, p.299.

Sung, Hyungun, and Ju-Taek Oh. "Transit-oriented development in a high-density city: Identifying its association with transit ridership in Seoul, Korea."Cities 28, no. 1 (2011), pp. 70-82.

Szell, M. et al., 2012. Understanding mobility in a social petri dish. Scientific reports, 2, p.457.

Taylor, Brian D., and Camille NY Fink. "The factors influencing transit ridership: A review and analysis of the ridership literature." (2003).

Taylor, Brian D., Douglas Miller, Hiroyuki Iseki, and Camille Fink. "Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas." Transportation Research Part A: Policy and Practice 43, no. 1 (2009), pp. 60-77.

Tobin, J., 1969. A general equilibrium approach to monetary theory. Journal of money, credit and banking, 1(1), pp.15–29.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: Journal of the Econometric Society, pp.307–333.

Washington, S.P., Karlaftis, M.G. & Mannering, F.L., 2010. Statistical and econometric methods for transportation data analysis, CRC press.

Yang, H. et al., 2010. Nonlinear pricing of taxi services. Transportation Research Part A: Policy and Practice, 44(5), pp.337–348.

Yano, K., Nakaya, T. & Ishikawa, Y., 2000. An analysis of inter-municipal migration flows in Japan using GIS and spatial interaction modeling. GEOGRAPHICAL REVIEW OF JAPAN SERIES B, 73(2), pp.165–177.

Yuan, J. et al., 2011. Where to find my next passenger. In Proceedings of the 13th international conference on Ubiquitous computing. pp. 109–118.

Zhao, F., Chow, L. F., Li, M. T., Ubaka, I., & Gan, A. (2003). Forecasting transit walk accessibility: regression model alternative to buffer method. Transportation

Research Record: Journal of the Transportation Research Board, 1835(1), pp. 34-41.

APPENDIX

Estimation Results for ZINB Models

Table A.1 Estimation Results for Variables at Origins (Non-Zero State)

Variable		Weekday			Weekend		
		Morning	Off	Evening	Morning	Off	Evening
Constant	parameter	6.829	8.897	6.116	5.941	8.241	6.106
	t-stat	50.829	42.5	61.432	27.957	31.321	32.51
Trip Distance	parameter	-0.515	-0.803	-0.491	-0.37	-0.632	-0.447
	t-stat	-47.732	-24.844	-34.972	-20.564	-19.435	-22.389
	marginal	-0.110	-0.174	-0.104	-0.081	-0.134	-0.100
Black Population	parameter	-0.632	-0.77	-0.639	-0.532	-0.693	-0.7
	t-stat	-47.732	-52.11	-47.429	-28.073	-34.716	-32.935
	marginal	-0.135	-0.167	-0.136	-0.117	-0.147	-0.157
Number of jobs	parameter	0.869	0.945	0.792	0.8	0.925	0.923
	t-stat	46.533	36.879	42.373	29.054	31.684	29.938
	marginal	0.185	0.205	0.169	0.176	0.197	0.207
Unemployment rate	parameter		-0.61				
	t-stat		-3.911				
	marginal		-0.132				
Higher annual income (1- if the average annual income in a ZCTA is more than115,000, 0-otherwise)	parameter	-5.41	-7.55	-5.782	-5.041	-7.457	-7.415
	t-stat	-35.173	-40.863	-38.687	-21.536	-30.051	-28.203
	marginal	-1.153	-1.635	-1.230	-1.108	-1.587	-1.660
Land use mixture index	parameter		1.575	0.992		1.273	1.759
	t-stat		12.89	8.222		7.755	10.292
	marginal		0.341	0.211		0.271	0.394
Commuters' mean travel time (1- if the average commuters' travel time is more than 25 min, 0-otherwise)	parameter						
	t-stat						
	marginal						
Colleges (1- if a ZCTA with one or more colleges, 0-otherwise)	parameter	0.829	0.782	0.694	0.772	0.86	0.702
	t-stat	19.633	15.950	16.321	12.906	12.406	9.824
	marginal	0.177	0.169	0.148	0.170	0.183	0.157
Recreational sites (1-if a ZCTA with more than 3 recreational sites, 0-otherwise)	parameter		0.465	0.541		0.589	0.401
	t-stat		7.721	9.377		7.017	4.5
	marginal		0.101	0.115		0.125	0.090

Table A.2 Estimation Results for Variables at Destinations (Non-Zero State)

Variable		Weekday			Weekend		
		Morning	Off	Evening	Morning	Off	Evening
Black Population	parameter	-0.403	-0.416	-0.439	-0.313	-0.445	-0.43
	t-stat	-29.082	-31.252	-36.957	-17.398	-23.141	-21.564
	marginal	-0.086	-0.090	-0.093	-0.069	-0.095	-0.096
Number of jobs	parameter	0.458	0.659	0.83	0.473	0.811	0.873
	t-stat	18.936	27.835	49.575	15.1	25.532	30.131
	marginal	0.098	0.143	0.177	0.104	0.173	0.195
Unemployment rate	parameter	-1.703	-0.886		-0.962	-0.958	
	t-stat	-11.819	-5.743		-4.372	-4.02	
	marginal	-0.363	-0.192		-0.211	-0.204	
Higher annual income (1- if the average annual income in a ZCTA is more than115,000, 0-otherwise)	parameter	-4.854	-6.654	-6.31	-5.206	-8.118	-7.764
	t-stat	-27.562	-36.895	-43.51	-18.292	-28.478	-27.859
	marginal	-1.035	-1.441	-1.343	-1.144	-1.728	-1.738
Land use mixture index	parameter	2.103	1.092	0.383	0.91	1.559	1.532
	t-stat	14.9	9.002	3.638	4.829	9.24	8.907
	marginal	0.448	0.237	0.082	0.200	0.332	0.343
Commuters' mean travel time (1- if the average commuters' travel time is more than 25 min, 0-otherwise)	parameter	-2.856	-2.671	-2.884	-2.372	-2.769	-2.767
	t-stat	-68.995	-61.331	-78.840	-43.503	-47.143	-43.436
	marginal	-0.609	-0.579	-0.614	-0.521	-0.589	-0.620
Colleges (1- if a ZCTA with one or more colleges, 0-otherwise)	parameter	0.531	0.826	0.771	0.935	0.771	0.599
	t-stat	10.481	15.473	16.074	13.11	10.241	7.992
	marginal	0.113	0.179	0.164	0.205	0.164	0.134
Recreational sites (1-if a ZCTA with more than 3 recreational sites, 0-otherwise)	parameter	0.474					
	t-stat	6.689					
	marginal	0.101					

Table A.3 Estimation Results for Variables at Origins (Zero State)

Variable		Weekday			Weekend		
		Morning	Off	Evening	Morning	Off	Evening
Constant	parameter	-32.483	-13.008	-38.062	-16.92	-8.426	-8.532
	t-stat	-29.081	-29.279	-33.325	-15.887	-25.739	-21.013
Trip Distance	parameter	2.794	2.239	2.824	2.934	2.425	2.424
	t-stat	66.343	56.808	68.024	39.654	41.086	39.989
	marginal	0.596	0.485	0.601	0.645	0.516	0.543
Black Population	parameter						
	t-stat						
	marginal						
Number of jobs	parameter	-0.526		-0.411	-0.339		
	t-stat	-17.83		-14.017	-8.059		
	marginal	-0.112		-0.087	-0.074		
Unemployment rate	parameter	8.964	0.611	12.069		0.706	
	t-stat	26.594	5.153	34.609		3.881	
	marginal	1.911	0.132	2.568		0.150	
Higher annual income (1- if the average annual income in a ZCTA is more than115,000, 0-otherwise)	parameter	20.486		22.129	11.71		
	t-stat	20.068		20.951	10.28		
	marginal	4.368		4.709	2.573		
Land use mixture index	parameter	-0.983		-1.658	-2.831		
	t-stat	-6.004		-9.416	-10.749		
	marginal	-0.210		-0.353	-0.622		
Commuters' mean travel time (1- if the average commuters' travel time is more than 25 min, 0-otherwise)	parameter	2.135	1.655	2.71	2.765	1.993	1.732
	t-stat	42.779	36.941	47.125	31.923	29.915	23.547
	marginal	0.455	0.358	0.577	0.607	0.424	0.388
Colleges (1- if a ZCTA with one or more colleges, 0-otherwise)	parameter	-1.242	-1.676	-1.775	-0.975	-1.429	-1.294
	t-stat	-14.806	-19.149	-21.673	-7.633	-12.351	-11.848
	marginal	-0.265	-0.363	-0.378	-0.214	-0.304	-0.290
Recreational sites (1-if a ZCTA with more than 3 recreational sites, 0-otherwise)	parameter						
	t-stat						
	marginal						

Table A.4 Estimation Results for Variables at Destinations (Zero State)

Variable		Weekday			Weekend		
		Morning	Off	Evening	Morning	Off	Evening
Black Population	parameter						
	t-stat						
	marginal						
Number of jobs	parameter	-0.22	-0.391	-0.622	-0.18		
	t-stat	-6.477	-12.644	-20.894	-4.593		
	marginal	-0.047	-0.085	-0.132	-0.040		
Unemployment rate	parameter	6.559	4.846	7.093	1.883	2.186	3.03
	t-stat	17.341	15.007	21.471	4.859	7.287	8.89
	marginal	1.398	1.050	1.509	0.414	0.465	0.678
Higher annual income (1- if the average annual income in a ZCTA is more than115,000, 0-otherwise)	parameter	10.967	10.804	15.84	6.505	3.073	2.975
	t-stat	16.29	19.155	27.519	11.767	12.252	10.167
	marginal	2.338	2.340	3.371	1.429	0.654	0.666
Land use mixture index	parameter	-0.735			-1.089		
	t-stat	-4.116			-3.681		
	marginal	-0.157			-0.239		
Commuters' mean travel time (1- if the average commuters' travel time is more than 25 min, 0-otherwise)	parameter						
	t-stat						
	marginal						
Colleges (1- if a ZCTA with one or more colleges, 0-otherwise)	parameter	-1.111	-0.458	-0.671	-0.836	-0.584	-0.708
	t-stat	-14.67	-6.734	-9.013	-6.726	-5.829	-6.904
	marginal	-0.237	-0.099	-0.143	-0.184	-0.124	-0.159
Recreational sites (1-if a ZCTA with more than 3 recreational sites, 0-otherwise)	parameter	-0.672	-0.371		-0.915	-0.705	
	t-stat	-6.672	-4.351		-6.848	-5.976	
	marginal	-0.143	-0.080		-0.201	-0.150	